Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Matemáticas
Escuela Académico Profesional de Estadística

Estudio del índice de desarrollo humano de los países de América: una aplicación del análisis multivariado

MONOGRAFÍA

Para optar el Título Profesional de Licenciado en Estadística

AUTOR

Alfredo LEÓN AGUILAR

ASESOR

Violeta Alicia NOLBERTO SIFUENTES

Lima, Perú

2008

Referencia bibliográfica

León, A. (2008). *Estudio del índice de desarrollo humano de los países de América: una aplicación del análisis multivariado*. Monografía para optar el título profesional de Licenciado en Estadística. Escuela Académico Profesional de Estadística, Facultad de Ciencias Matemáticas, Universidad Nacional Mayor de San Marcos, Lima, Perú.

OBJETIVO DEL ESTUDIO

General

Caracterizar a los países del continente americano empleando variables usadas para construir el Indicador de Desarrollo Humano mediante el Análisis Multivariante.

Específicos

1.- Elaborar indicadores sintéticos empleando las variables usadas para construir el Indicador de Desarrollo Humano, mediante el Análisis Factorial.

2.- Identificar los países de América que tengan características similares empleando Análisis Cluster.

3.- Evaluar el agrupamiento de los países del continente americano usando las caras de Chernoff.

CAPÍTULO I

INDICADORES DE DESARROLLO HUMANO

1.1. INTRODUCCIÓN

La construcción de indicadores se inicia con la selección del fenómeno a estudiar, la recopilación de la información relacionada con dicho fenómeno, las variables que serán usadas para analizarlo, el modelo que relacione esas variables, el análisis y la interpretación de los resultados. El indicador puede responder a una razón simple de dos cantidades, una proporción, a un porcentaje, a una tasa o a un índice (Barrantes, 2006).

Por tanto se requiere definir adecuadamente la información del fenómeno a estudiar. Amartya Sen[1] define al desarrollo, como la expansión de las libertades reales de las que dispone una persona para hacer y ser lo que valora en la vida. El desarrollo es por tanto un proceso de liberación desde una situación de privaciones. Según Ceara – Hatton[2], dice que una persona sin educación, sin salud, desnutrida y sin empleo no puede elegir ni tiene opciones en la vida. De esta definición se desprenden tres implicancias:

- La primera, que el desarrollo es una cuestión de poder porque las libertades reales dependen del acceso a las oportunidades y en sociedades con un deficiente estado de derecho y con una gran inequidad social, como en América, el acceso a las oportunidades depende más del poder personal y de grupos y no tanto de los derechos. Aún más; para crear una sociedad con estado de derecho hay que asegurar institucionalmente que toda persona pueda ejercer su poder.

[1] Economista bengalí, conocido por sus trabajos sobre las hambrunas, la teoría del desarrollo humano, la economía del bienestar y los mecanismos subyacentes de la pobreza. Recibió el Premio Nobel de Economía en 1998.
[2] Director General del informe sobre Desarrollo Humano (República Dominicana, 2008).

Sólo así se asegura el ejercicio efectivo de los derechos de toda la población.

- La segunda, que el desarrollo tiene una dimensión individual y otra colectiva. Las capacidades, es decir, las habilidades y condiciones personales para poder hacer o ser lo que se desea en la vida, se portan individualmente, pero se construyen socialmente. En efecto, las posibilidades de tener educación de calidad o de tener acceso a un sistema de justicia efectivo dependen de circunstancias institucionales que van más allá de las posibilidades.
- La tercera, se asocia al hecho que quienes se desarrollan son las del individuo. Pero además, las personas viven en sociedad, por tanto la calidad y la forma de las relaciones interpersonales forman parte del modelo de bienestar de las personas, no las cosas ni la macroeconomía, ni la modernidad. Las personas viven en una realidad y en un espacio concreto, establecen un conjunto de relaciones y vínculos de cooperación, de poder y de afectos. Por lo tanto, el desarrollo también está enlazado con el lugar en dónde vive la gente.

1.2. DEFINICIÓN DE VARIABLE

Según el Diccionario de la lengua española[3] una variable es: "Medida de una cantidad con referencia a una escala determinada."

Según W. Mendenhall[4], una variable es una característica que cambia o varía con el tiempo y/o para los diferentes individuos u objetos que se consideren. En su acepción general, variable se refiere a una característica que se mide en el estudio, no son fijas, pueden cambiar e intervienen muchos factores en este cambio. En términos estadísticos rigurosos, una variable es la representante de esas mediciones en el análisis. Según Mendenhall existen 2 tipos de variables:

1. Cualitativas o categóricas
2. Cuantitativas
    a. Discretas

[3] Diccionario Lengua Española, 2001, p.1074.
[4] Mendenhall, W., Wackerly, D., y Scheafer, R. (2002). Estadística Matemática con Aplicaciones. México: Thompson.

b. Continuas

Por otro lado, según Ruiz Muñoz y Sánchez Sánchez[5], una variable aleatoria es una función que asigna un valor numérico a cada suceso elemental del espacio muestral. Es decir, una variable aleatoria es aquella cuyo valor numérico está determinado por el resultado del experimento aleatorio. La variable aleatoria la notaremos con letras en mayúscula X, Y, ... y con las letras en minúscula x, y, ... sus valores. La variable aleatoria puede ser numerable o no numerable y además de ser discreta o continua.

1.3. DEFINICIÓN DE INDICADOR

Un indicador es una variable simple utilizada en conjunción con una o más variables para formar una medición compuesta. Elemento o atributo que mide un proceso o producto.

El proceso de "gestión" de datos como tarea político-administrativa y de investigación supone el reconocimiento previo de un complejo proceso de construcción de la información (Cinta de Mohebio, "El Proceso de Gestión de Datos, 2002)".

Desde esta perspectiva, resulta fundamental entender que los datos no están "dados en la realidad" y que solo resta recopilarlos, sino que son fruto de una acción creadora y por tanto condicionada por las perspectivas teórico-metodológicas desde las cuales se los construye. Además a esta afirmación la "gestión de información" debe recurrir a datos que deben ser susceptibles de ser contados, medidos, pesados y para ello se apoya en una metodología cuantitativa que se basa en diversas fuentes. Los indicadores como elementos fundamentales de este paradigma La medida, construcción de índices y de indicadores, de manejo estadístico de masas más o menos grandes de datos carecen en general de fronteras. Y aquí lo que nos interesa resaltar es el importante tema de la construcción de indicadores sociales, que surge ante la necesidad de cuantificar determinadas dimensiones de una situación social,

[5] David Ruiz Muñoz y Ana María Sánchez Sánchez: (2006) Apuntes de Estadística, Edición electrónica: http://www.eumed.net/libros/2006a/rmss/a7.htm , consulta el 30 de junio de 2008.

como por ejemplo la satisfacción ante una determinada prestación social o el nivel de vida de una determinada población. En este caso, los indicadores son utilizados como puros instrumentos de conocimiento del fenómeno bajo estudio. Diríamos como el básico instrumento de conocimiento, sin discutir de dónde provienen o cuál es la base teórica de su origen. Al respecto cabe destacar el importante aporte de Blalock[6] quien sostiene que no existe una correspondencia directa entre teoría y realidad, entre conceptos y observaciones, por lo que "se requiere la existencia de una 'teoría auxiliar' como intermediaria entre ambos planos, que especifique en cada caso el modo de relación de un indicador determinado con una variable teórica determinada". Esta preocupación que planteamos subyace en todo sistema de información que funciona como "gestión de datos" sin problematizar la estandarización de esquemas conceptuales y teóricos que tienden a homogeneizar la información en una progresiva organización burocrática de la misma. La organización internacional de sistemas de indicadores omite la discusión crítica de enfoques teóricos y metodológicos y se constituye en recetas de planificación y evaluación de proyectos y programas sociales. Esta paradigmática postura ha llevado a la primacía de la producción masiva de datos.

1.4. IMPORTANCIA DEL DESARROLLO HUMANO

Todo país en vías de desarrollo debe incentivar planes educacionales, sociales, culturales, de la salud que favorezcan a quienes viven en el territorio sin discriminación de clases, ofrecerles servicio más proactivos con calidad, excelencia, no con improvisación y discriminación, que favorezca a todos (PNUD, 2007).

El Programa de las Naciones Unidas para el Desarrollo (PNUD) señala, que el bienestar de las personas no depende únicamente de la cantidad de bienes que se tiene o recibe –como lo concebiría una visión económica del desarrollo, sino también de las necesidades satisfechas en lo referente al acceso a la salud, la educación y el respeto de sus derechos fundamentales. Estas variables conforman el concepto de Desarrollo Humano (DH), cuyo

[6] Blalock, H. 1968 Methodology in social research, Nueva York, McGraw-Hill.

origen se debió en gran parte a la colaboración de los economistas Mahbub ul Hag[7] y Amartya Sen[8].

El PNUD define el Desarrollo Humano en su informe de 1990 como *"El proceso en el que se amplían las oportunidades del ser humano. En principio, estas oportunidades pueden ser infinitas y cambiar con el tiempo. Sin embargo, de todos los niveles del desarrollo, son tres las más esenciales: disfrutar de una vida prolongada y saludable, adquirir conocimientos y tener acceso a los recursos necesarios para lograr un nivel de vida decente".*

Resalta en este concepto que *"la verdadera riqueza de una nación está en su gente. El objetivo básico del desarrollo es crear un ambiente propicio para que los seres humanos disfruten de una vida prolongada, saludable y creativa".* Según esta concepción, *"invertir en la gente es la mejor estrategia de crecimiento económico: una sociedad progresa sólo si todas las personas que la integran incrementan sus libertades y capacidades, y las instituciones se concentran en trabajar para la gente y con la gente"*. El proceso de Desarrollo Humano incluye varios aspectos de la interacción humana como la participación, la equidad de género, la seguridad, la sostenibilidad, la garantía de los derechos humanos y otros aspectos que son reconocidos por las personas como imprescindibles.

El concepto de desarrollo humano surgió y tomó fuerza en la última década del siglo XX como una reacción crítica frente al déficit y limitaciones del concepto que en mayor medida se venia utilizando a escala mundial desde fines de la Segunda Guerra Mundial, como era el concepto de desarrollo económico.

Asimismo, desde el año 1990, el PNUD está impulsando la definición de Desarrollo Humano a través de la publicación anual de un informe. En éste se plasma cada año la situación mundial frente a diferentes temas relacionados al Desarrollo Humano, los cuales se han incorporado a los criterios e indicadores que permiten medir y analizar el proceso de desarrollo de los países. Y es que

[7] Mahbub ul Haq, economista pakistaní fundador del Informe sobre Desarrollo Humano, fallecido en 1998

[8] Amartya Sen, Premio Nobel de Economía 1998

un proceso tan complejo como lo es el del Desarrollo Humano necesita ser abordado de diferentes ópticas y atender las diferentes realidades que tienen los países. No obstante, esta práctica fue adoptada también por diversos Estados, quienes elaboran informes nacionales y locales como instrumentos de monitoreo para observar el progreso o fijar metas para alcanzar mejores niveles en los indicadores de desarrollo. Generalmente los Institutos de Estadística de cada país, -que es el órgano rector en materia de estadísticas económicas, sociales y demográficas- construyen y publican periódicamente el Índice de Desarrollo Humano, con la colaboración metodológica del PNUD.

Esta estrategia dio como resultado la producción de un conjunto de orientaciones conceptuales y de herramientas metodológicas dirigidas a facilitar la puesta en marcha del Desarrollo Humano Sostenible Local (DHSL). La estrategia del DHSL constituye una forma sistemática de intervenir en las comunidades pobres ampliando las oportunidades de las personas mediante la capacitación y la organización para el desarrollo local, fomentando simultáneamente la concertación de los diversos agentes de desarrollo para el aprovechamiento de las fortalezas de las condiciones sociales, culturales, económicas, ambientales y políticas de los territorios. La orientación de esta estrategia la proporciona el concepto de Desarrollo Humano, definido como el proceso de ampliación de las opciones de las personas con el fin de que la gente lleve la vida que considera valiosa. El PNUD se ha planteado construir alianzas para asesorar y capacitar en la estrategia de DHSL. El objetivo es fortalecer las capacidades de hombres y mujeres, organizaciones e instituciones, y promover oportunidades de desarrollo social, económico, político, institucional, territorial, ambiental y cultural.

1.5. INDICADOR DE DESARROLLO HUMANO (IDH)

El IDH es una medida sinóptica[9] del desarrollo humano. Mide el progreso medio conseguido por un país en tres dimensiones básicas del desarrollo humano:

[9] Una medida sinóptica está elaborada por un conjunto de medidas de diferentes variables

1.- Disfrutar de una vida larga y saludable, medida a través de la esperanza de vida al nacer.
2.- Disponer de educación, medida a través de la tasa de alfabetización de adultos (con una ponderación de dos terceras partes) y la tasa bruta combinada de matriculación en primaria, secundaria y terciaria (con una ponderación de una tercera parte).
3.- Disfrutar de un nivel de vida digno, medido a través del PIB per cápita en términos de la paridad del poder adquisitivo (PPA) en dólares estadounidenses (US$).

1.6. MEDICIÓN DE DESARROLLO HUMANO

Según López-Calva y Vélez[10], las ponderaciones adecuadas para el cálculo del IDH han sido centrales en las discusiones académicas y todas las propuestas de indicadores multidimensionales se enfrentan a ellas. El PNUD ha utilizado para el cálculo del IDH tres dimensiones básicas: ingreso, educación y salud. Cada dimensión es medida a través de un índice que considera las desviaciones o "rezagos" con respecto a algún valor de referencia, que en general es el logro deseable. Una vez obtenidos los índices en cada dimensión, éstos se agregan con ponderaciones iguales[11].

Existe mucha literatura donde se pueden encontrar críticas al IDH, fundamentalmente en dos niveles: el conceptual y el de implementación. Las críticas conceptuales se basan en la discusión sobre las dimensiones relevantes a considerar y su forma de agregación.

En este sentido, Ravallion (1997) y Kelley (1991) señalan que la forma de agregación del IDH por medio de un promedio simple hace posible un nivel perfecto de sustitución entre sus tres componentes. Ravallion (1997) hace énfasis en que los juicios detrás de la valoración de las distintas dimensiones del IDH no son explícitos y, por lo tanto, son cuestionables. Kelley (1991), por su parte, observa que el PNUD no presenta un análisis detallado que justifique el uso de la fórmula de agregación del IDH; él sugiere que se dé un peso mayor

[10] El concepto de desarrollo humano, su importancia y aplicación en México, Luis F. Lopez-Calva y Roberto Velez, 2003, PNUD- México
[11] Para una explicación detallada de la metodología de cálculo del IDH ver el Apéndice.

al PIB per-capita, ya que el ingreso refleja la capacidad de los individuos de escoger entre muchas otras dimensiones del desarrollo humano. Sin embargo, aceptar esta propuesta es acercarse de nuevo al enfoque del ingreso y alejarse del enfoque de las capacidades.

Este último autor somete a discusión algunos problemas teóricos de la medición del IDH. En primer término, critica la definición de las cotas basadas en valores "ideales" de las variables consideradas, que se utilizan para construir los índices de cada componente. Para ilustrar, el punto cambia el valor superior o "cota deseada" de la esperanza de vida de 78 a 73 años y encuentra que 22 países pasarían de la categoría de "desarrollo humano bajo" a la de "desarrollo humano medio" y que otros 10 países pasarían de "desarrollo humano medio" a "desarrollo humano alto[12]".

Por último, Kelley hace notar que la metodología de medición del IDH no permite reflejar el mejoramiento de los países catalogados como desarrollados porque su índice presenta valores que van de 0.96 a 1. Srinivasan (1994) alude a dos cuestiones en lo que al marco conceptual se refiere. La primera tiene que ver con el origen del concepto. En su opinión, el uso del enfoque "alternativo" data desde 1938, año en el que el Indian National Planning Comite enfatizó la necesidad de lograr mejores niveles de vida para la población a través de la reducción de la pobreza y de la búsqueda de una distribución más equitativa de la riqueza.

Srinivasan defiende la idea de que el rol instrumental del ingreso ha estado presente desde hace tiempo en el pensamiento de los economistas y sobre todo de los planeadores de política pública. Por otra parte, este autor afirma que no hay correspondencia entre el concepto de las capacidades y su implementación a través del IDH debido a que las capacidades no tienen una métrica homogénea, a diferencia del enfoque del ingreso que sí tiene esa característica. Srinivasan señala en este sentido que, dado que el valor relativo

[12] El PNUD clasifica a los países con índices mayores a 0.8 como de "desarrollo humano alto", a los países con un índice entre 0.5 y 0.799 como de "desarrollo humano medio", y a los países con índices menores a 0.5como de "desarrollo humano bajo".

de las capacidades varía entre individuos, grupos sociales y países y, tomando en cuenta que los componentes del IDH son funcionamientos en el sentido de Amartya Sen, el valor intrínseco de cada funcionamiento no se captura en la medición lineal del IDH.

Respecto a la implementación del cálculo del IDH, el autor en cuestión observa que pueden surgir ciertos problemas derivados de la misma construcción de las bases de datos que se utilizan para hacer los cálculos del índice. Además de los posibles sesgos que lo anterior puede generar, señala que el IDH tiene serios problemas para ser comparable en el tiempo y en el espacio principalmente debido a las variaciones que ha sufrido su metodología de cálculo, lo que dificulta las inferencias acerca de los procesos de desarrollo.

Si bien las críticas anteriores tienen su grado de validez, es cierto también que la importancia y la utilidad del IDH han sido validadas empíricamente. Hoy en día, los sectores académicos y de política pública aceptan que el IDH provee información relevante sobre el estado del desarrollo desde una perspectiva amplia y los índices complementarios.

1.7. USO DEL INDICADOR DE DESARROLLO HUMANO

La novedad principal del IDH es la creación de una estadística única que sirva como marco de referencia tanto para el desarrollo social como para el económico. Es así que la utilización del IDH como indicador del nivel de desarrollo de una población se basa en la idea, generalmente aceptada hoy en los medios políticos y académicos, de que si bien el crecimiento económico es una condición necesaria para explicar el grado de avance de un país, no constituye una condición suficiente. Por otro lado, la utilización del IDH ha sido mejorada al desagregarlo. Por ejemplo, es posible que el índice general para un país oculte las diferencias en los niveles de desarrollo humano entre distintos grupos. Los IDH desagregados[13] se construyen con los datos de los componentes del IDH para cada grupo por separado. Los grupos pueden definirse según el ingreso, regiones geográficas o administrativas, la residencia

[13] Se denomina así cuando se analiza por cada uno de sus componentes.

urbana o rural, el género y la etnia. El uso de IDH desagregados en los ámbitos nacionales y subnacionales ayuda a destacar las desigualdades y deficiencias importantes entre regiones, géneros, áreas rurales y urbanas y grupos étnicos. El análisis que los IDH desagregados hacen posible debería ayudar a orientar las políticas y las acciones para atacar las desigualdades y deficiencias.

Es posible que las desigualdades ya sean muy conocidas, pero el IDH las puede revelar de manera aún más nítida. El IDH desagregado por grupos sociales o regiones también posibilita que los grupos comunitarios locales soliciten más recursos y obliguen a los representantes locales a rendir cuentas. Esto hace del IDH una herramienta para el desarrollo participativo.

Desde que fueron creados, los IDH desagregados se han utilizado en gran medida para el análisis de países como Brasil, Colombia, Egipto, Gabón, Alemania, India, Kazajstán, Malasia, México, Nigeria, Papua Nueva Guinea, Polonia, Sudáfrica, Trinidad y Tobago, Turquía, Ucrania y los EE. UU. Los Informes Nacionales sobre Desarrollo Humano de China y Kenya encontraron grandes desigualdades entre provincias y entre regiones urbanas y rurales. Un estudio similar en Guatemala demostró que esas desigualdades se aplican además a grupos étnicos.

CAPÍTULO II

TÉCNICAS MULTIVARIANTES DE INTERDEPENDENCIA

2.1. INTRODUCCIÓN

En el presente capítulo trataremos de dos técnicas multivariantes de interdependencia, el Análisis Factorial y el Análisis Cluster, métodos estadísticos multivariantes que se usa en el presente trabajo. Se describe el modelo factorial y el Análisis Cluster, las razones de su uso y propiedades importantes. También se explica la evaluación de la matriz de correlaciones, los métodos de estimación para este caso, el número de factores a extraer, la interpretación de factores y la rotación de los factores para el Análisis Factorial, y finalmente la interpretación de los grupos obtenidos, empleando Análisis Cluster.

2.2. ANÁLISIS FACTORIAL

El análisis factorial es un nombre genérico que se da a una clase de métodos estadísticos multivariantes cuyo propósito principal es definir la estructura subyacente en una matriz de datos. Aborda el problema de cómo analizar la estructura de las correlaciones entre un gran número de variables (Hair J.F., 1999).

Con el análisis factorial se puede identificar primero las dimensiones separadas de la estructura de la matriz de correlaciones de variables que describen algún concepto y determinar el grado en que se justifica cada variable por cada dimensión. Una vez que se determinan estas dimensiones y se explican las relevancias de cada variable, se pueden lograr los dos objetivos

principales para el análisis factorial que son: el resumen y la reducción de datos.

El resumen de datos, con el análisis factorial permiten obtener dimensiones subyacentes que, cuando son interpretadas y comprendidas, describen los datos con un número de conceptos mucho más reducido que las variables individuales originales. Se puede obtener la reducción de datos con el cálculo de la puntuación para cada dimensión subyacente y sustituirlos por los valores de las variables originales (Hair J.F., 1999)

Si el número de variables es demasiado grande o existe necesidad de dar una mejor representación a un número de conceptos más pequeños, el análisis factorial puede asistir en la selección de un subgrupo de variables representativo o incluso crear nuevas variables llamadas factores (que no son observables directamente), como sustitutas para las variables originales mientras mantengan su carácter original.

2.2.1. OBJETIVOS DEL ANÁLISIS FACTORIAL

El principal objetivo de análisis factorial es analizar la estructura de la matriz de covarianzas o de correlaciones, es decir, describir si es posible las relaciones entre varias variables en términos de pocas variables no observables pero importantes llamados Factores, de modo que expliquen una buena parte de la variabilidad de las varianzas de cada variable original (Johnson,1982, C. M. Cuadras).

2.2.2. MODELO FACTORIAL

El modelo matemático del Análisis Factorial es parecido al de la regresión múltiple. Cada variable se expresa como una combinación lineal de factores no directamente observables (Johnson, 1982).

Sean $X_1, X_2, ..., X_p$ variables bajo estudio; tal que $\vec{X} \sim \left(\vec{0}, \Sigma\right)$ el modelo básico se expresa como:

$$X_1 = \mu_1 + l_{11}F_1 + l_{12}F_2 + .. + l_{1m}F_m + \varepsilon_1$$

$$X_2 = \mu_2 + l_{21}F_1 + l_{22}F_2 + .. + l_{2m}F_m + \varepsilon_2$$

. . . .

. . . . .

. . . .

$$X_p = \mu_p + l_{p1}F_1 + l_{p2}F_2 + .. + l_{pm}F_m + \varepsilon_p \quad \text{.....(II.1)}$$

2.2.2.1 Supuestos del modelo factorial

1.- $m < p$ puesto que se desea explicar las variables por un número más reducido de variables teóricas llamadas factores.

2.- La totalidad de los $(m + p)$ factores son incorrelacionados; se pretende que la parte de la variabilidad de una variable explicada por un factor no tenga relación (en sentido lineal) con los demás factores.

La expresión (II.1) puede expresarse matricialmente como:

$$\vec{X} = \vec{\mu} + L\vec{F} + \vec{\varepsilon} \quad \text{.....(II.2)}$$

Donde:

$\vec{X}^T = (X_1, X_2, ..., X_p)$: Es el vector de variables observadas

$\vec{\mu}^T = (\mu_1, \mu_2, ..., \mu_p)$: Denota al vector de medias poblacionales.

$\vec{F}^T = (F_1, F_2, .., F_m)$: Es un vector de m variables no observables, llamado factores comunes; porque de acuerdo al modelo (II.1) influyen en común en las p variables.

$\vec{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_{p)}$: Es un vector de p variables no observables llamado vector de factores específicos o vector de factores únicos.

$L = (l_{ij})$: Es una matriz de dimensión (pxm), que contiene constantes desconocidas, llamadas cargas factoriales o pesos factoriales.

2.2.3. SUPUESTOS DEL MODELO FACTORIAL ORTOGONAL

Tres son los supuestos de este modelo:

1.- $E(\vec{F}) = \underset{(mx1)}{0}$, $Cov(\vec{F}) = E(\vec{F}\vec{F}^T) = \underset{(mxm)}{I}$,

2.- $E(\vec{\varepsilon}) = \underset{(px1)}{0}$, $Cov(\vec{\varepsilon}) = E(\vec{\varepsilon}\vec{\varepsilon}^T) = \underset{(pxp)}{\Psi}$;donde Ψ es una matriz diagonal.

3.- $Cov(\vec{\varepsilon}, \vec{F}) = E(\vec{\varepsilon}, \vec{F}^T) = 0_{(pxm)}$ .....(II.3)

Cabe recalcar que los factores únicos no están correlacionados entre si ni con los factores comunes, la matriz de covarianza de los factores específicos se expresa como:

$$\Psi = \begin{bmatrix} \Psi & 0 & 0 & 0 \\ 0 & \Psi_2 & 0 & 0 \\ . & . & . & . \\ 0 & 0 & 0 & \Psi_p \end{bmatrix} \quad \text{..... (II.4)}$$

2.2.4. DESCOMPOSICIÓN DE LA MATRIZ DE COVARIANZAS SEGÚN EL MODELO FACTORIAL

Para determinar como es la estructura de la matriz de covarianza Σ, usaremos los supuestos del modelo (II.3), llamado modelo factorial ortogonal y la definición de covarianza, bajo operaciones y propiedades anteriormente mencionadas obtendremos el siguiente resultado:

$$\Sigma = Cov(\vec{X}) = E[(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^T]$$

$$= E[(L\vec{F} + \vec{\varepsilon})((L\vec{F})^T + \vec{\varepsilon})]$$

$$= E[(L\vec{F}(L\vec{F})^T) + \vec{\varepsilon}(L\vec{F})^T + L\vec{F}\vec{\varepsilon}^T + \vec{\varepsilon}\vec{\varepsilon}^T]$$

$$= LE(\vec{F}\vec{F}^T)L^T + E(\vec{\varepsilon}\vec{F}^T)L^T + LE(\vec{F}\vec{\varepsilon}^T) + E(\vec{\varepsilon}\vec{\varepsilon}^T)$$

$$\Sigma = LL^T + \Psi \qquad \text{……. (II.5)}$$

La matriz Σ queda descompuesta en función de $L_{(pxm)}$, matriz de pesos del vector $\vec{F}$ y Ψ la matriz de varianzas específicas de $\vec{X}$.

La varianza de X_i y la covarianza de $X_{i,}X_j$ se expresa como:

$$Var(X_i) = \sigma_{ii} = {l_{i1}}^2 + {l_{12}}^2 + ... + {l_{im}}^2 + \Psi_i \qquad \text{…….(II.6)}$$

$$Cov(X_{i,}X_k) = \sigma_{ik} = l_{i1}l_{k1} + l_{i2}l_{k2} + ... + l_{im}l_{km} \qquad \text{…….(II.7)}$$

Donde: $h_i^2 = {l_{i1}}^2 + {l_{i2}}^2 + ... + {l_{im}}^2$

es la i-ésima comunalidad, que indica el grado de asociación que tiene esta variable original a través del factor i-ésimo con el resto de variables; es la variabilidad compartida de las variables a través del factor común.

Del resultado (II.6) se observa que la varianza de una variable tiene dos componentes, bajo el modelo factorial ortogonal, esto es:

$$\sigma_{ii} = h_i^2 + \Psi_i \rightarrow \quad \sigma_{ii} = comunalidad + \text{var}\,ianza.específica \qquad \text{……. (II.8)}$$

La varianza específica es aquella variabilidad que está asociada con la variable específica.

$Cov(X,F) = L$, esto es $Cov(X_i, F_j) = l_{ij}$ ……. (II.9)

Si el análisis factorial es adecuado para explicar las interrelaciones o interdependencias entre variables, la comunalidad debe ser lo más alto posible

en comparación con la varianza específica. Ahora bien, si se factoriza la matriz de correlaciones, la diagonal está dada por:

$$1 = comunalidad + \text{var}ianza\ específica$$

2.2.5. ESQUEMA PARA UN ADECUADO ANALISIS FACTORIAL

El esquema que se suele seguir para un adecuado Análisis Factorial es:

1. Realizar el cálculo de la matriz de correlaciones entre todas las variables, conocida habitualmente como matriz R.

2. A partir de la matriz de correlaciones, extraer los factores necesarios para representar los datos.

3. Si es necesario, realizar la rotación de los factores con el objeto de facilitar su interpretación.

4. Calcular las puntuaciones factoriales de cada individuo.

En realidad sólo los dos primeros pasos son indispensables, el 3º y 4º son un complemento, dependerá de los objetivos que se pretenden lograr en una investigación.

2.2.6. EXAMEN DE LA MATRIZ DE CORRELACIONES

El primer paso en el Análisis Factorial es calcular la matriz de correlaciones considerando todas las variables que entran en el análisis. Un Análisis Factorial resultará adecuado cuando existan altas correlaciones entre las variables, y permite suponer que se explican por factores comunes (Vinacua, 1998).

Pueden utilizarse diferentes métodos para comprobar el grado de asociación entre las variables y verificar la adecuación del modelo entre ellas tenemos:

2.2.6.1. El Determinante de la Matriz de Correlaciones:

Un determinante muy bajo indicará altas intercorrelaciones entre las variables o lineal dependencia, pero no debe ser cero (matriz no singular), pues

esto indicaría que algunas de las variables son linealmente dependientes y no se podrían realizar ciertos cálculos necesarios en el Análisis Factorial (Vinacua, 1998).

2.2.6.2. Índice Kaiser- Meyer - Olkin (KMO)

Es una medida de adecuación del modelo factorial, el KMO se calcula como:

$$KMO = \frac{\sum\sum_{i \neq j} r_{ij}^{\ 2}}{\sum\sum_{i \neq j} r_{ij}^{\ 2} + \sum\sum_{i \neq j} a_{ij}^{\ 2}} \quad \text{……. (II.10)}$$

Donde:

r_{ij} : es el coeficiente de correlación de Pearson entre la variables i y j.

a_{ij} : es el coeficiente de correlación parcial entre las variables i y j

Estos coeficientes compara los coeficientes de correlación de Pearson con los coeficientes de correlación entre variables.

Si los coeficientes de correlación entre las variables son muy pequeños, esto quiere decir que la relación entre cada para de las mismas se debe o puede ser explicada por el resto y por lo tanto llevar a cabo un análisis factorial de los datos no deja de ser una buena solución.

En este supuesto, si la suma de los coeficientes de correlación parcial al cuadrado es muy pequeña, el KMO será un índice muy próximo a la unidad y por tanto el análisis factorial un procedimiento adecuado. En cambio valores pequeños en este índice nos da a entender todo lo contrario.

Para la evaluación de la adecuación del modelo factorial y su interpretación, se emplea los siguientes niveles:

$1 \geq$ KMO > 0.90 Excelentes

$0.90 \geq$ KMO > 0.80 Buenos

$0.80 \geq$ KMO > 0.70 Aceptables

$0.70 \geq$ KMO > 0.60 Mediocres o regulares

$0.60 \geq$ KMO > 0.50 Malos

KMO ≤ 0.50 Inaceptables o muy malos.

2.2.6.3. Test de Bartlett

Se utiliza para verificar si la matriz de correlaciones es una matriz identidad, es decir si todos los coeficientes de la diagonal son iguales a la unidad y los extremos a la diagonal iguales a 0; que indica ausencia de correlación entre pares de variables distintas.

Este estadístico se obtiene a partir de la transformación χ^2 del determinante de la matriz de correlaciones y cuanto mayor sea el valor χ^2 y menor el grado de significación más improbable que la matriz sea una matriz identidad. En el supuesto de que no se pudiese rechazar esta hipótesis, no se debe realizar el análisis factorial con los datos.

La prueba de esfericidad de Bartlett propone las siguientes hipótesis estadísticas:

$$H_o : \Sigma = I$$
$$H_1 : \Sigma \neq I$$

Si, como resultado del contraste, rechazamos la Ho, entonces podemos realizar el Análisis Factorial, ya que las variables están correlacionados, se emplea solo si la muestra multivariante es multinormal (Vinacua, 1998).

2.2.6.4. Medida de Adecuación Muestral (MSA)

La medida de adecuación muestral (MSA), son los coeficientes que se encuentran en la diagonal de la matriz de correlación anti – imagen (se emplea

solo si la muestra multivariante es multinormal), que vienen a ser los KMO pero en este caso para cada variable por separado.

Esta medida está dada por:

$$MSA(j) = \frac{\sum_{i \neq j} r_{ij}^{\ 2}}{\sum_{i \neq j} r_{ij}^{\ 2} + \sum_{i \neq j} a_{ij}^{\ 2}} \quad \text{.....(II.11)}$$

Si el valor MSA fuera pequeño, no se aconsejaría el Análisis Factorial, por el contrario, valores próximos a 1 indicarían que la variable X_i es adecuada para incluirla con el resto en un Análisis Factorial. En muchas ocasiones, se eliminan las variables con MSA muy bajo (Vinacua, 1998).

2.2.7. MÉTODO DE ESTIMACIÓN

Una vez que se especifican las variables y se prepara la matriz de correlación, ya se puede aplicar el análisis factorial. Para ello es necesario definir el método de estimación de los factores.

Entre los métodos de estimación de factores, tenemos al método de Componentes Principales y Máxima Verosimilitud (R. Johnson, 1982), que se describen a continuación.

2.2.7.1. Método de Componentes Principales

Este método también es conocido como el método de Factores Principales, se basa en suponer que los factores comunes explican el comportamiento de las variables originales en su totalidad, usa la descomposición espectral de la matriz de covarianzas así entonces:

$\sum$ por la descomposición espectral contiene los pares $(\lambda_i, \vec{\gamma}_i)$ con $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$,donde λ_i son los valores propios (autovalores) y $\vec{\gamma}_i$ son los vectores propios (autovectores) generados a partir de los λ_i entonces :

$$\Sigma = \lambda_1 \vec{\gamma}_1 \vec{\gamma}_1^{\ T} + \lambda_2 \vec{\gamma}_2 \vec{\gamma}_2^{\ T} + ... + \lambda_p \vec{\gamma}_p \vec{\gamma}_p^{\ T}$$

$$\left[\sqrt{\lambda_1}\vec{\gamma}_1 \quad \sqrt{\lambda_2}\vec{\gamma}_2 \quad .....\sqrt{\lambda_p}\,\vec{\gamma}_p\right]\begin{bmatrix}\sqrt{\lambda_1}\vec{\gamma} \\ \sqrt{\lambda_2}\vec{\gamma}_2 \\ . \\ .. \\ \sqrt{\lambda_p}\vec{\gamma}_p\end{bmatrix} \quad \text{.....(II.12)}$$

La estructura de Σ, tiene tantos factores como variables bajo estudio y la varianza específica es cero para todas las variables, obviamente que la matriz de factores tiene como j-ésima columna, $\sqrt{\lambda_j}\vec{\gamma}_j$ por tantoΣ se puede escribir como:

$$\Sigma = LL^T + 0 = LL^T \quad \text{.....(II.13)}$$

Pero como buscamos m factores comunes, entonces p-m factores no se consideran debido a que p-m autovalores son despreciables, por tanto:

$$\Sigma = \left[\sqrt{\lambda_1}\vec{\gamma}_1 \quad \sqrt{\lambda_2}\vec{\gamma}_2 \quad .....\sqrt{\lambda_p}\,\vec{\gamma}_p\right]\begin{bmatrix}\sqrt{\lambda_1}\vec{\gamma} \\ \sqrt{\lambda_2}\vec{\gamma}_2 \\ . \\ .. \\ \sqrt{\lambda_p}\vec{\gamma}_p\end{bmatrix} = L_{(pxm)}L_{(mxp)}{}^T \quad \text{.....(II.14)}$$

donde $L_{(pxm)}$ es la matriz de pesos factoriales.

La representación aproximada en (II.14) asume que el factor específico $\vec{\varepsilon}$ es de menor importancia y puede ignorarse en la factorización de Σ, si los factores específicos son incluidos en el modelo, sus varianzas pueden obtenerse de: $\Sigma - LL^T$, donde LL^T se define en (II.14), entonces la aproximación es:

$$\Sigma = LL^T + \Psi \quad \text{.....(II.15)}$$

Donde $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}{}^2$ para $i = 1,2,...,p$ .....(II.16)

Ordenando y resumiendo lo anteriormente tratado, para la obtención del modelo factorial ortogonal mediante el método de componentes principales, para la matriz de covarianzas muestral S, la matriz de pesos factoriales es:

$$\hat{L} = \left[\sqrt{\hat{\lambda}_1}\hat{\gamma}_1 \quad \sqrt{\hat{\lambda}_2}\hat{\gamma}_2 \quad .....\sqrt{\hat{\lambda}_m}\hat{\gamma}_m\right] \quad \text{.....(II.17)}$$

Las estimaciones de las varianzas específicas se obtienen de la diagonal de la matriz:

S - $\hat{L}\hat{L}^T$, de manera tal que:

$$\hat{\Psi} = \begin{bmatrix} \hat{\Psi}_1 & 0 & 0 & 0 \\ 0 & \hat{\Psi}_2 & 0 & 0 \\ . & . & . & . \\ 0 & 0 & 0 & \hat{\Psi}_p \end{bmatrix} \quad \text{.....(II.18)}$$

$\hat{\hat{\psi}}_i = S_{ii} - \sum_{j=1}^{m} l_{ij}^{\,2}$ y las comunalidades estimadas son:

con

$$\hat{h}_i^{\,2} = \hat{l}_{i1}^{\,2} + \hat{l}_{i2}^{\,2} + ... + \hat{l}_{im}^{\,2}$$

.....(II.19)

En caso se obtenga el modelo factorial ortogonal mediante la matriz de correlaciones muestrales R, es necesario reemplazar adecuadamente a fin de obtener las comunalidades y las varianzas específicas correspondientes.

Este método de estimación es de mayor uso debido a que no hay condición alguna acerca de la distribución de probabilidad de la población, de donde se ha tomado la muestra aleatoria; además cuando el investigador desea considerar más factores de las que seleccionó, puede incrementarlos y la solución no cambia (A. Johnson, 1982).

2.2.8. NÚMERO DE FACTORES

Cuando un gran número de variables originales se analizan para la extracción de factores, el método realiza la extracción de combinaciones lineales de las variables que explican la cantidad mayor de la varianza y después continúa con combinaciones que justifican cantidades de varianza cada vez menores. Para decidir cuántos factores se deben estimar se empieza

generalmente con algún criterio predeterminado, tal como el porcentaje de varianza o el criterio de raíz latente, para llegar a un número de factores específicos.

Después de estimar la solución inicial, se calculan varias soluciones de prueba adicionales, normalmente un factor menos que el número inicial y dos o tres factores más que los que se estimaron inicialmente. Posteriormente, en función de la información que se obtiene de estos análisis previos, se examinan las matrices de factores y se escoge el número de factores que represente mejor los datos (Hair J.F., 1999).

No obstante, en general se utilizan los siguientes criterios para determinar el número de factores a considerar.

2.2.8.1. Criterio de Kaiser

Este criterio se utiliza más frecuentemente por la sencillez en su aplicación cuando se descompone la matriz de correlaciones. La racionalidad que se usa para el criterio de raíz latente o autovalor es que cualquier factor individual debería justificar la varianza de por lo menos una única variable.

Cada variable contribuye con un valor de 1 para el autovalor total. Por lo tanto, sólo se consideran los factores que tienen raíces latentes o autovalores mayores que 1; explican al menos una variable, se considera que todos los factores con raíces latentes menores que 1 (explican menos de 1 variable) no son significativas y por tanto, se desestiman a la hora de incorporarlos a la interpretación.

El uso del autovalor para establecer una corte es más fiable cuando el número de variables está entre 20 y 50. Si el número de variables es menor de 20, hay una tendencia a que este método extraiga un número conservador de factores; por el contrario, si hay más de 50 variables, lo más común es que se extraigan demasiados factores (Hair J.F, 1999).

2.2.8.2. Criterio a priori

El criterio a priori es un criterio simple y a la vez razonable bajo ciertas circunstancias. Su aplicación, nos permite saber cuantos factores hay que extraer antes de iniciar el análisis factorial. Consiste en instruir al computador para detener el análisis cuando se haya extraído el número de factores deseado. Esta aproximación resulta de utilidad cuando se prueba una teoría o una hipótesis acerca del número de factores para ser extraído. Este criterio depende mucho de la experiencia del investigador (Hair J.F, 1999).

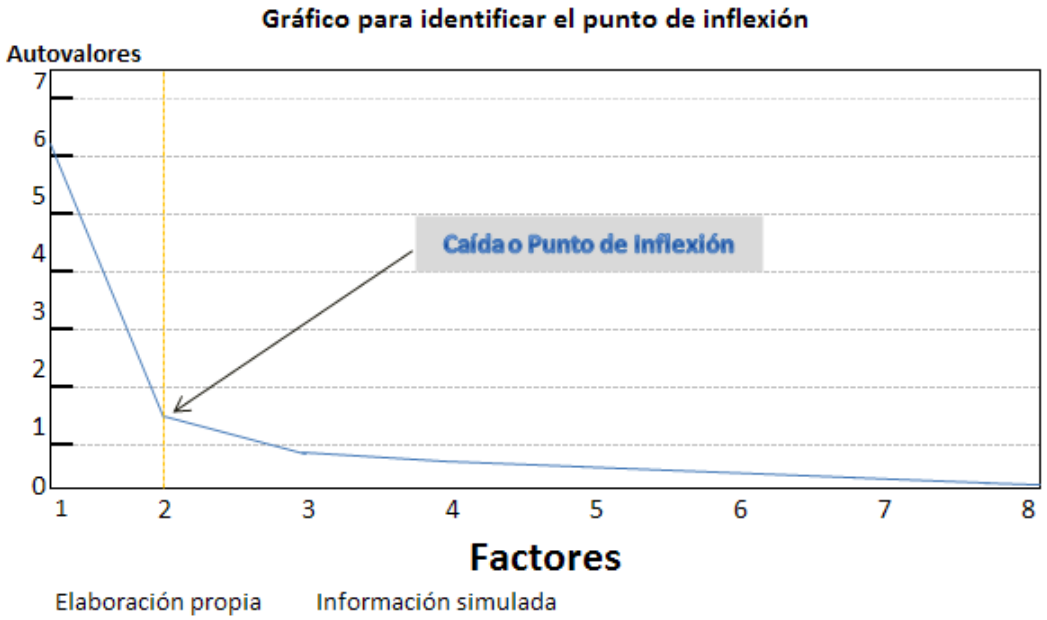
2.2.8.3. Criterio de porcentaje de la varianza

El criterio de porcentaje de la varianza acumulada respecto a la varianza total equivale ha obtener un porcentaje acumulado especificado de la varianza total extraída. El propósito de este criterio es asegurar una significación práctica de los factores derivados, asegurando que explican por lo menos una cantidad especificada de la varianza. No se ha adoptado un umbral absoluto para todas las aplicaciones. Sin embargo en las ciencias naturales, el procedimiento de factores no debería ser detenido hasta que los factores extraídos cuenten por lo menos con un 95% de la varianza acumulada. Por contraste en las ciencias sociales, donde la información muchas veces es menos precisa, es normal considerar una solución que represente un 60% de la varianza total como satisfactoria (Hair J.F, 1999).

2.2.8.4. Criterio de contraste de caída (Catell)

El contraste de caída se utiliza para identificar el número óptimo de factores que pueden ser extraídos antes de que la cantidad de la varianza única empiece a dominar la estructura de la varianza común. Se estima el contraste de caída con el trazo de autovalores en función del número de factores en su orden de extracción, y se utiliza la forma de curva consiguiente para evaluar el punto de corte (Hair J.F, 1999).

En la siguiente figura, se presenta el gráfico de “codos” o de la pendiente de “Cattell”, en el que se observa que al presentar el autovalor de cada uno de los 8 factores, el quiebre o inflexión se muestra en el autovalor del factor 2, por tanto se sugiere solo 2 factores.



Elaboración propia Información simulada

2.2.9. INTERPRETACIÓN DE LOS FACTORES

Para interpretar los factores y seleccionar la solución factorial definitiva, en primer lugar se calcula la matriz inicial no rotada, para tener una indicación preliminar acerca del número de factores a extraer.

La matriz de factores contiene los pesos factoriales para cada variable original sobre cada factor. Al calcular la matriz de factores no rotada, nos interesa la mejor combinación lineal de variables originales, es decir encontrar aquella combinación particular de variables originales que cuenta con el mayor porcentaje de varianza de los datos. En consecuencia, el primer factor puede contemplarse como el mejor resumen de las relaciones lineales que los datos manifiestan. El segundo factor se define como la segunda combinación lineal de las variables, sujeta a la restricción de que sea ortogonal al primer factor. Para se ortogonal al primer factor, el segundo factor debe derivarse de la varianza restante tras la extracción del primer factor.

Así, el segundo factor puede definirse como la combinación lineal de las variables que da cuenta del mayor porcentaje de varianza residual una vez se ha eliminado de los datos el efecto del primer factor. Los factores subsiguientes se definen de manera análoga hasta haber agotado la varianza de los datos y los últimos factores tienen menor porcentaje de varianza.

Las soluciones factoriales no rotadas alcanzan el objetivo de reducción de datos, pero pueden ser que no faciliten la interpretación más adecuada de las variables examinados. La solución factorial no rotada puede no suministrar un patrón significativo de los pesos de las variables, si se espera que los factores no rotados sean significativos, el usuario puede especificar que la rotación no se lleva a cabo.

Generalmente, la rotación es deseable porque simplifica la estructura de factores, y facilita la interpretación, también habitualmente es difícil determinar si los factores no rotados serán relevantes (Hair J.F, 1999).

2.2.10. ROTACIÓN DE FACTORES

La obtención, inicialmente por una transformación ortogonal, de los pesos factoriales, reproducen la estructura de la matriz de covarianzas o correlaciones, pero cuando no podemos interpretar el sentido y significado de los factores, se recurre a la rotación hasta obtener una estructura simple y darle la interpretación debida a los factores comunes.

Es importante la rotación de factores para su interpretación, consiste en girar el origen de los ejes de referencia de los factores hasta alcanzar una determinada posición. Las soluciones factoriales no rotadas extraen factores según su orden de importancia. El primer factor tiene mayor porcentaje de varianza, el segundo y siguientes factores se basan en la varianza residual. Cada uno explica porcentajes de varianza cada vez menores. El efecto último de rotar la matriz de factores es redistribuir la varianza de los primeros factores a los últimos para lograr un patrón de factores más simples y teóricamente más significativo (Hair J.F, 1999).

Johnson (1982) enuncia, si $\hat{L}$ es la matriz estimada ^{pxm} de pesos factoriales obtenida por el método de componentes principales, entonces $\hat{L}^* = \hat{L}T$; donde $TT^T = T^T T = I$, es una matriz pxm de pesos rotados. Además la matriz de covarianza estimada (o matriz de correlación) permanece inalterada dado que:

$$\hat{L}\hat{L}^T + \hat{\Psi} = \hat{L}TT^T\hat{L} + \hat{\Psi} = \hat{L}^*\hat{L}^{*T} + \Psi \qquad \text{.....(II.21)}$$

la ecuación anterior indica que la matriz residual $S_n - \hat{L}\hat{L}^T - \hat{\Psi} = S_n - \hat{L}^*\hat{L}^{*T} - \hat{\Psi}$ no cambia. Además las varianzas específicas $\hat{\Psi}$ y las comunalidades no se alteran. Dado que los pesos factoriales originales no son interpretables en la práctica es usual rotarlas hasta que se obtenga una estructura más simple (Johnson, 1982).

Existen diversos tipos de rotación, los ortogonales (Varimax, Cuartimax y Equamax, entre otros), y los no ortogonales (Promax, Quartimin, Oblimin entre otros), de todos estos métodos el más utilizado es el Varimax. (Affifi y Clark, Visauta 1998). Se describe dos de cada tipo:

2.2.10.1. Rotación Varimax

El método Varimax se utiliza con mayor frecuencia. Fue diseñado por Kaiser (1958) y, como su nombre lo indica está orientado a la maximización de la varianza de factores. Lógicamente, una forma de conseguir esto es que algunos factores tengan pesos altos y otros próximos a cero, lo que tiene como consecuencia que bastantes variables tengan puntuaciones bajas en un factor.

Por trigonometría elemental, sabemos que para rotar un eje tenemos que multiplicar por coseno y seno del ángulo de rotación, de modo que debemos primero identificar el ángulo que deseamos rotar los ejes. Kaiser (1958) propuso un criterio de seleccionar que es empleado por paquetes estadísticos como SPSS. Concretamente, el ángulo de rotación seleccionado j, será aquel que minimice la expresión siguiente:

$$V = \frac{\sum_{k=1}^{c}\sum_{j=1}^{p} g^4{}_{kj} - (\sum_{k=1}^{c}\sum_{j=1}^{p} g^2{}_{kj})^2}{p^2} \qquad \text{.....(II.22)}$$

2.2.10.2. Cuartimax

Trata de minimizar el número de factores necesarios para explicar un conjunto de variables. El objetivo de este tipo de rotación es simplificar las filas de una matriz de factores; cuartimax se centra en rotar los factores iniciales de tal forma que una variable cargue alto sobre un factor y tan bajo como sea

posible sobre los otros factores. En estas rotaciones muchas variables pueden cargar alto o cerca sobre el mismo factor que la técnica se centra en las filas (Visauta, 1998).

2.2.10.3. Promax

La rotación Promax es otro método directo que ha tenido notable aceptación. Introducido por Hendrickson y White (1964), se caracteriza por ser directamente calculable, sin necesidad de procesos iterativos, siendo mucho más simple de cálculo que los demás métodos de rotación. Promax se aplica directamente a la matriz factorial ortogonal, según el criterio varimax. Para obtener una estructura mas simple, bastará destacar los factores haciendo crecer las saturaciones altas y decrecer las saturaciones bajas (C. M. Cuadras, 1981). Esta rotación es aconsejable cuando se trabaja con grandes bases de datos (trabaja con parámetro λ cuyo valor por defecto es 4).

2.2.10.4. Oblimin

Trabaja con el parámetro δ. El valor de δ será 0 si queremos que los factores sean muy oblicuos, o lo que es lo mismo que exista una alta correlación entre los mismos y tanto más negativo su valor cuanto más oblicuos (Visauta, 1998).

2.2.11. PUNTUACIONES FACTORIALES

En el análisis factorial interesa usualmente centrarse en los parámetros del modelo factorial. Sin embargo, los valores estimados de los factores comunes, llamado puntaje factoriales (scores factor), pueden también ser requeridos. Estas cantidades son después usadas para propósito de diagnóstico como buenas salidas de análisis posteriores.

Los puntajes factoriales no son estimaciones de parámetros desconocidos en el sentido usual. Por el contrario ellos son valores estimados para el vector de factores aleatorios no observados.

$F_j, j = 1,2,3...n$ Esto es puntaje factorial

$\hat{f}_j =, j = 1,2,3...n$ Estimación del valor

$\hat{f}_j$ obtenido por F_j (j ésimo caso)

Uno de los principales objetivos del Análisis Factorial es reducir un gran número de variables a un pequeño número de factores. Es aconsejable estimar las puntuaciones de las unidades de análisis con la finalidad de interpretar y categorizarlos.

Existen diversos métodos para obtener las puntuaciones factoriales, como el método de Regresión, el método de suma de cuadrados ponderados entre otros (Jhonson, 1982, Visauta, 1998).

2.3. ANÁLISIS CLUSTER

El Análisis Cluster es un conjunto de técnicas que se utilizan para clasificar los objetos o casos en grupos relativamente homogéneos llamados conglomerados (clusters). Los objetos en cada grupo (conglomerado) tienden a ser similares entre sí (alta homogeneidad interna, dentro del cluster) y diferentes a los objetos de los otros grupos (alta heterogeneidad externa, ente clusters) con respecto a algún criterio de selección predeterminado. De este modo, si la clasificación es un éxito, los objetos dentro del cluster estarán muy cercanos unos de otros en la representación geométrica, y los clusters diferentes estarán muy apartados (Jhonson, 1982, Visauta, 1998).

El Análisis Cluster tiene como propósito esencial, agrupar aquellos objetos que reúnan idénticas características, es decir, se convierte así en una técnica de análisis exploratorio diseñada para revelar las agrupaciones naturales dentro de una colección de datos. Este análisis no hace ninguna distinción entre variables dependientes (VD) y variables independientes (VI) sino que calcula las relaciones interdependientes de todo el conjunto de variables. Es decir, el Análisis Cluster puede utilizarse como instrumento de reducción general de datos a fin de desarrollar subgrupos de datos que sean más fáciles de manejar que las observaciones individuales.

Con el Análisis Cluster, no debe esperarse nunca una única y definitiva solución al problema de creación de conglomerados, de ahí que es exploratorio. En la práctica, se considerarán varias soluciones derivadas de la

aplicación de las muchas técnicas ofrecidas por este análisis y, tras su correspondiente comparación, nos inclinaremos por la más útil para el tipo de investigación planteada. Los conocimientos del investigador sobre el problema, jugarán un papel importante a la hora de decidir entre las diferentes soluciones alternativas.

2.3.1. PLANTEAMIENTO DEL PROBLEMA

Sean $X_1,...,X_p$ p variables cuantitativas observadas en n objetos.
Sea x_{ij} = valor de la variable X_j en el i-ésimo objeto i=1,...,n; j=1,...,p.

El objetivo del Análisis Cluster es obtener grupos de objetos de forma que, por un lado, los objetos pertenecientes a un mismo grupo sean muy semejantes entre sí, es decir, que el grupo esté cohesionado internamente y, por el otro, los objetos pertenecientes a grupos diferentes tengan un comportamiento distinto con respecto a las variables analizadas, es decir, que cada grupo esté aislado externamente de los demás grupos.

Es una técnica eminentemente exploratoria puesto que la mayor parte de las veces, no utiliza ningún tipo de modelo estadístico para llevar a cabo el proceso de clasificación. Es una técnica muy adecuada para extraer información de un conjunto de datos sin imponer restricciones previas en forma de modelos estadísticos, al menos de forma explícita y, por ello, puede llegar a ser muy útil como una herramienta de elaboración de hipótesis acerca del problema considerado sin imponer patrones o teorías previamente establecidas.

2.3.2. MEDIDAS DE SIMILITUD

Se sabe que el objetivo del Análisis Cluster o Conglomerado es agrupar objetos similares; una vez establecidas las variables y los objetos o unidades a clasificar el siguiente paso consiste en establecer una medida de proximidad o de distancia entre ellos, que cuantifique el grado de similaridad entre cada par de objetos (Hair. J.F, 1999).

El concepto de similaridad es fundamental en el Análisis Cluster. La similaridad (similitud) es una medida de correspondencia o semejanza entre los objetos que van a ser agrupados.

La similitud entre objetos puede medirse de varias formas, entre ellos tenemos, las medidas de proximidad y las medidas distancia. Las medidas de proximidad y de distancia dependen del tipo de variables y datos considerados.

2.3.2.1. Clase de datos

a) Intervalar: se trata de una matriz objetos x variables en donde todas las variables son cuantitativas, medidas en escala intervalo o razón.

b) Frecuencias: las variables analizadas son categóricas de forma que, por filas, tenemos objetos o categorías de objetos y, por columnas, las variables con sus diferentes categorías. En el interior de la tabla aparecen frecuencias.

c) Datos binarios: se trata de una matriz objetos x variables pero en la que las variables analizadas son binarias de forma que 0 indica la ausencia de una característica y 1 su presencia.

2.3.3. MEDIDAS DE PROXIMIDAD

Las medidas de proximidad, también son llamadas medidas de similitud o semejanza, miden el grado de semejanza entre dos objetos de forma que, cuanto mayor es su valor, mayor es el grado de similaridad existente entre ellos y con más probabilidad los métodos de clasificación tenderán a ponerlos en el mismo grupo. Son importantes las consideraciones que incluyen la naturaleza de la variable (discreta, continua y binaria) o medidas en escala (nominal, ordinal, intervalar o razón) y sujeto a materia de conocimiento.

2.3.3.1. Medidas para variables cuantitativas

a).- Coeficiente de congruencia

$$c_{rs} = \frac{\sum_{j=1}^{p} x_{rj} x_{sj}}{\sqrt{\sum_{j=1}^{p} x_{rj}^2}\sqrt{\sum_{j=1}^{p} x_{sj}^2}} \qquad \text{.....(II.23)}$$

que es el coseno del ángulo que forman los vectores $(x_{r1},...,x_{rp})'$ y $(x_{s1},...,x_{sp})'$.

b).- Coeficiente de correlación

$$r_{rs} = \frac{\sum_{j=1}^{p} \left(x_{rj} - \overline{x}_r\right)\left(x_{sj} - \overline{x}_s\right)}{\sqrt{\sum_{j=1} \left(x_{rj} - \overline{x}_r\right)^2}\sqrt{\sum_{j=1}^{p} \left(x_{sj} - \overline{x}_s\right)^2}} \qquad \text{.....(II.24)}$$

donde $\overline{x}_r = \frac{\sum_{j=1}^{p} x_{rj}}{p}$ y $\overline{x}_s = \frac{\sum_{j=1}^{p} x_{sj}}{p}$ . .....(II.25)

Si los objetos r y s son variables, r_{rs} mide el grado de asociación lineal existente entre ambas.

Estas dos medidas se utilizan, preferentemente, para clasificar variables siendo, en este caso, invariantes por cambios de escala y, en el caso del coeficiente de correlación, invariante por cambio de origen. Por esta razón es más conveniente utilizar el coeficiente de congruencia con variables medidas bajo escala de razón[14] en las cuales el origen está claramente definido.

Conviene observar, además, que tanto el coeficiente c_{rs} como r_{rs} toman valores comprendidos entre -1 y 1 pudiendo tomar, por lo tanto, valores negativos. Dado que en algunos casos (por ejemplo, si los objetos a clasificar son variables), los valores negativos cercanos a -1 pueden implicar fuerte

[14] Las variables con escala de tipo de razón poseen las mismas características de las variables de intervalo, con la diferencia que cuentan con un cero absoluto; es decir, el valor cero (0) representa la ausencia total de medida, por lo que se puede realizar cualquier operación *Aritmética.*

semejanza entre los objetos clasificados, conviene en estas situaciones, utilizar como medida de semejanza sus valores absolutos.

2.3.3.2. Medidas para datos binarios

En este caso se construyen, para cada par de objetos **r** y **s**, tablas de contingencia de la forma:

Objeto r \ Objeto s	0	1
0	a	b
1	c	d

Donde:

0 indica ausencia

1 indica presencia

Además, a = número de variables en las que los objetos **r** y **s** toman el valor 0, etc. y p = a+b+c+d. Utilizando dichas tablas algunas de las medidas de semejanza más utilizadas entre otros:

Coeficiente de Jaccard: $\frac{d}{b+c+d}$ .....(II.26)

Coeficiente de acuerdo simple: $\frac{a+d}{p}$ .....(II.27)

Ambas toman valores entre 0 y 1 y miden, en tanto por uno, el porcentaje de acuerdo en los valores tomados en las p variables, existente entre los dos objetos. Difieren en el papel dado a los acuerdos en 0. El coeficiente de Jaccard no los tiene en cuenta y el de acuerdo simple si los considera. Ello es debido a que, en algunas situaciones, las variables binarias consideradas son asimétricas en el sentido de que es más informativo el valor 1 que el valor 0. Además, de los coeficientes de similitud antes descritos, existen otros que se usan de acuerdo a la necesidad del estudio, presentamos algunos.

Otras medidas de similitud (Variables binarias)

Coeficiente	Descripción
$\frac{2(a+d)}{2(a+d)+b+c}$	Doble peso a coincidencias 0-0 y 1-1
$\frac{a+d}{a+d+2(b+c)}$	Doble peso a las no coincidencias
$\frac{d}{p}$	No se tienen en cuenta las coicidencias 0-0 en el numerador
$\frac{a}{a+b+c}$	No se tienen en cuenta las coincidencias 1-1
$\frac{2d}{2d+b+c}$	No se tienen en cuenta las coincidencias 0-0, y se da doble peso a las coincidencias 1-1
$\frac{d}{d+2(b+c)}$	No se tienen en cuenta las coincidencias 0-0, y se da doble peso a las no coincidencias
$\frac{d}{b+c}$	Cociente de coincidencias y no coincidencias, con exclusión de coincidencias 0-0

2.3.4. MEDIDAS DE DISTANCIA

La medidas de distancia también llamada medida de disimilitud y desemejanza, miden la distancia entre dos objetos de forma que, cuanto mayor sea su valor, más diferentes son los objetos y menor la probabilidad de que los métodos de clasificación los pongan en el mismo grupo.

2.3.4.1. Medidas para variables cuantitativas

Se usan con mayor frecuencia las siguientes:

La *distancia euclídea* y la *distancia euclídea al cuadrado*, están dadas por las siguientes notaciones respectivamente:

$$\sqrt{\sum_{j=1}^{p}\left(x_{rj}-x_{sj}\right)^2} \quad \text{y} \quad \sum_{j=1}^{p}\left(x_{rj}-x_{sj}\right)^2 \qquad \text{.....(II.28)}$$

La distancia euclídea al cuadrado, nos dice que los conglomerados o clusters deben constituirse de tal manera que, al juntarse dos elementos, la perdida de información resultante debe ser mínima. Por lo tanto la cantidad de

información se cuantifica como la suma de las distancias al cuadrado de cada elemento respecto al elemento del centro del conglomerado al que pertenece.

La distancia métrica de Chebychev: $\max_i |x_{ri}-x_{si}|$

Distancia de Manhattan: $\sum_{i=1}^{p} \left| X_{ri} - X_{si} \right|$ .....(II.29)

Distancia de Minkowski: $\sqrt[q]{\sum_{i=1}^{p} \left(X_{ri} - X_{si} \right)^q}$ con $q \in N$. .....(II.30)

Las tres primeras medidas son variantes de la distancia de Minkowski con q=2, ∞ y 1, respectivamente. Cuanto mayor es q más énfasis se le da a las diferencias en cada variable.

Todas estas distancias no son invariantes a cambios de escala por lo que se aconseja estandarizar los datos si las unidades de medida de las variables no son comparables. La forma más común de estandarización es la conversión de cada variable en puntuaciones típicas (también conocidas como puntuaciones Z). La forma de cálculo es que se resta a cada observación de cada variable su media correspondiente y se divide el resultado de esta operación por la desviación típica (estándar) de la variable en cuestión. Este proceso convierte la puntuación de cada dato original en un valor estandarizado con una media de cero y una desviación típica de uno. (Hair J.F., 1999)

En definitiva, lo que se consigue con ello es eliminar, uno por uno, los prejuicios introducidos por las diferencias en las escalas de los distintos atributos (variables) usados en el análisis. Además, estas distancias no consideran las relaciones existentes entre las variables. Si se requieren considerarlos se aconseja utilizar la distancia de *Mahalanobis* que viene dada por la forma cuadrática:

$$(\bar{X}_r - \bar{X}_s)' S^{-1} (\bar{X}_r - \bar{X}_s) \quad \text{.....(II.31)}$$

donde $\vec{x}_r = (x_{r1},...,x_{rp})'$ y $\vec{x}_s = (x_{s1},...,x_{sp})'$

2.3.4.2. Medidas para tablas de frecuencias

Están basadas en la χ^2 de Pearson. Algunas de las más utilizadas son:

$$\chi^2 = \sqrt{\sum_{i=1}^{p} \frac{(x_{ri} - E(x_{ri}))^2}{E(x_{ri})} + \sum_{i=1}^{p} \frac{(x_{si} - E(x_{si}))^2}{E(x_{si})}} \quad \text{.....(II.32)}$$

$$\chi^2 = \sqrt{\frac{\sum_{i=1}^{p} \frac{(x_{ri} - E(x_{ri}))^2}{E(x_{ri})} + \sum_{i=1}^{p} \frac{(x_{si} - E(x_{si}))^2}{E(x_{si})}}{N}} \quad \text{.....(II.33)}$$

Donde $E(x_{ri}) = \frac{x_{r.} x_{.i}}{N}$ con $x_{r.} = \sum_{i=1}^{p} x_{ri}$ y $x_{.i} = x_{ri} + x_{si}$ .....(II.34)

representa el valor esperado de la frecuencia x_{ri} si hay independencia entre los individuos **r** y **s** y las categorías 1,...,p de las variables y $N = x_{r.} + x_{s.}$ es el total de observaciones. La diferencia entre ambas medidas radica en la división por N en el caso de χ^2 para paliar la dependencia que tiene la χ^2 de Pearson respecto a N.

2.3.4.3. Medidas para datos binarios

Al igual que en las medidas de proximidad o similitud, para la medida de distancia con datos binarios, se construye, para cada par de objetos r y s, tablas de contingencia de la forma:

Objeto r \ Objeto s	0	1
0	a	b
1	c	d

Donde:

0 indica ausencia

1 indica presencia

Donde a = número de variables en las que los objetos r y s toman el valor 0, etc. Utilizando dichas tablas se construye las medidas de distancia siendo las más utilizadas:

Distancia euclídea al cuadrado: $(DE)^2 = a + b$ .....(II.35)

Lance y Williams: $\frac{b+c}{2d+b+c}$ …….(II.36)

Esta última ignora las coincidencias en 0.

Otras medidas de distancia (Variables binarias)

Coeficiente	Descripción
$\frac{(b-c)^2}{p^2}$	Diferencia de tamaño
$\frac{bc}{p^2}$	Diferencia de configuración
$\frac{b+c}{4p}$	Varianza

2.3.4.4. Medidas para datos de tipo mixto

Si en la base de datos existen diferentes tipos de variables: binarias, categóricas, ordinales, cuantitativas, no existe una solución universal al problema de cómo combinarlas para construir una medida de distancia. Anderberg (1973) o Gordon (1990) sugieren las siguientes soluciones:

a.- Expresar todas las variables en una escala común, habitualmente binaria, transformando el problema en uno de los ya contemplados anteriormente. Esto tiene sus costes, sin embargo, en términos de pérdida de información si se utilizan escalas menos informativas como las nominales ú ordinales o la necesidad de incorporar información extra si se utilizan escalas más informativas como son el intervalo o la razón.

b.- Combinar medidas con pesos de ponderación mediante expresiones de la forma:

$$d_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} d_{ijk}}{\sum_{k=1}^{p} w_{ijk}} \qquad \text{…….(II.37)}$$

Donde d_{ijk} es la distancia entre los objetos i y j en la k-ésima variable y w_{ijk} = 0 ó 1 dependiendo de si la comparación entre i y j es válida en la k-ésima variable.

c.- Realizar análisis por separado utilizando variables del mismo tipo y utilizar el resto de las variables como instrumentos para interpretar los resultados obtenidos.

2.3.5. MÉTODOS DE CLASIFICACIÓN O SELECCIÒN

Existen dos tipos de procedimiento: los jerárquicos y los no jerárquicos.

2.3.5.1. Métodos jerárquicos

Este tipo de conglomerado se caracteriza por el desarrollo de una jerarquía o estructura en forma de árbol. Los métodos jerárquicos pueden ser por Aglomeración o por División. El análisis por aglomeración empieza considerando a cada objeto como un conglomerado individual. Los conglomerados se forman al agrupar los objetos en conjuntos cada vez más grandes. Este proceso continúa hasta que todos los objetos formen parte de un solo grupo. El conglomerado por división comienza con todos los objetos agrupados en un solo conjunto. Los conglomerados se dividen hasta que cada objeto sea un grupo independiente. En cada paso del algoritmo se divide el grupo más heterogéneo.

Luego, para establecer qué grupos se unen o dividen se utiliza una función objetivo o criterio que, en el caso de los métodos aglomerativos recibe el nombre de *enlace o eslabonamiento.*

- Tipos de enlace o eslabonamiento

Se emplean con los métodos aglomerativos y proporcionan diversos criterios para determinar, en cada paso del algoritmo, que grupos se deben unir. Podemos destacar los siguientes:

a) Enlace simple o vecino más próximo (single linkage)

También es conocido como método de enlace o eslabonamiento simple (Florek et al., 1951a, 1951b), mide la proximidad entre dos grupos calculando la distancia entre sus objetos más próximos o la similitud entre sus objetos más semejantes. Este método se basa en la distancia mínima, encuentra dos objetos separados por la distancia más corta y la coloca en el primer conglomerado. Luego encuentra la distancia más corta, y o bien un tercer objeto se une a los primeros para formar un conglomerado o se forma un nuevo conglomerado de dos miembros. El proceso continúa hasta que todos los objetos se encuentran en un conglomerado (Hair J.F., 1999).

b) Enlace promedio (average linkage)

Determina la proximidad entre dos grupos calculando la media de las distancias entre objetos de ambos grupos o la media de las similitudes entre objetos de ambos grupos. Así, por ejemplo, si se utilizan distancias, la distancia entre los grupos r y s vendría dada por:

$$\frac{1}{n_r n_s} \sum_{j \in r} \sum_{k \in s} d(j,k) \qquad \text{.....(II.38)}$$

Donde d(j,k) = distancia entre los objetos j y k y nr, ns son los tamaños de los grupos r y s, respectivamente. El método tiende a producir grupos de igual varianza (Milligan, 1980).

c) Enlace promedio ponderado (método de McQuitty)

Conocido también como método de McQuitty fue introducido independientemente por Sokal y Michener (1958) y McQuitty (1966). Representa una generalización del procedimiento anterior usando el número de objetos en cada conglomerado como peso. Vale decir que la distancia se basa en un promedio ponderado. Si los pesos son iguales este método otorga los mismos resultados que el método anterior.

d) Enlace medio dentro de los grupos

Mide la proximidad entre dos grupos con la distancia media existente entre los miembros del grupo unión de los dos grupos. Así, por ejemplo, si se trata de distancias, la distancia entre los grupos r y s vendría dada por:

$$\frac{1}{C^2_{n_r+n_s}} \sum_{(j,k)\in r\cup s} d(j,k) \qquad \text{.....(II.39)}$$

e) Métodos del centroide y de la mediana

Los dos métodos miden la proximidad entre dos grupos calculando la distancia entre sus centroides

$$d^2_{rs} = \sum_{j=1}^{p} \left(\overline{x}_{rj} - \overline{x}_{sj}\right)^2 \qquad \text{.....(II.40)}$$

Donde X_{rj} y X_{sj} son las medias de la variable X_j en los grupos r y s, respectivamente.

Ambos métodos difieren en la forma de calcular los centroides: el método del centroide utiliza las medias de todas las variables de forma que las coordenadas del centroide del grupo r = s ∪ t vendrán dadas por:

$$\overline{x}_{rj} = \frac{1}{n_r} \sum_{m=1}^{n_r} x_{rjm} = \frac{n_s}{n_s + n_t} \overline{x}_{sj} + \frac{n_t}{n_s + n_t} \overline{x}_{tj} \qquad \text{.....(II.41)}$$

j = 1,…,p

En el método de la mediana el nuevo centroide es la media de los centroides de los grupos que se unen

$$X_{rj} = \frac{1}{2} X_{sj} + \frac{1}{2} X_{tj} \qquad \text{.....(II.42)}$$

f) Método de Ward

El atributo más distintivo del método de Ward es que no emplea las distancias entre conglomerados para realizar la aproximación, sino que trata de hacer mínima la variabilidad entre conglomerados, esto es de hacer que cada conglomerado sea lo más homogéneo posible. La homogeneidad se mide mediante la suma de cuadrados de diferencias entre los sujetos dentro de un conglomerado. Por tanto se agruparán los dos sujetos más semejante, es decir cuya suma de cuadrados del error sea menor (Catena, 2003).

El método busca minimizar $\sum_r SSW_r$ donde SSW_r es para cada grupo r, las sumas de cuadrados intragrupo que viene dada por:

$$SSW_r = \sum_{m=1}^{n_r} \sum_{j=1}^{p} (x_{rjm} - \bar{x}_{rj})^2 \qquad \text{.....(II.43)}$$

Donde x_{rjm} denota el valor de la variable X_j en el m-ésimo elemento del grupo r. En cada paso del algoritmo une los grupos r y s que minimizan:

$$SSW_t - SSW_r - SSW_s = \frac{n_r n_s}{n_r + n_s} d_{rs}^{\ 2} \quad \text{.....(II.44)}$$

con t = r $\cup$s y $d_{rs}^{\ 2}$ la distancia entre los centroides de r y s.

2.3.5.2. Métodos no jerárquicos

El segundo tipo de procedimientos de conglomerados, los métodos de conglomerados no jerárquicos, con frecuencia se conocen como Agrupación de K Medias. Estos tipos de métodos son convenientes utilizarlos cuando los datos a clasificar son muchos y/o para refinar una clasificación obtenida utilizando un método jerárquico. Supone que el número de grupos es conocido a priori.

Existen varias formas de implementarlo pero todas ellas siguen, básicamente, los siguientes pasos:

1) Se seleccionan k centroides o semillas donde k es el número de grupos deseado.
2) Se asigna cada observación al grupo cuya semilla es la más cercana.
3) Se calculan los puntos semillas o centroides de cada grupo.
4) Se iteran los pasos 2) y 3) hasta que se satisfaga un criterio de parada como, por ejemplo, los puntos semillas apenas cambian o los grupos obtenidos en dos iteraciones consecutivas son los mismos. El método suele ser muy sensible a la solución inicial dada por lo que es conveniente utilizar una que

sea buena. Una forma de construirla es mediante una clasificación obtenida por un algoritmo jerárquico.

2.3.6. SELECCIÓN DEL NÚMERO DE CONGLOMERADOS

Se desconoce un criterio universalmente aceptado, porque se trata de un método exploratorio. Sin embargo, ya que la mayoría de los programas estadísticos proporciona las distancias de aglomeración, es decir, las distancias a las que se forma cada grupo, una forma de determinar el número de grupos consiste en localizar en qué iteraciones del método utilizado dichas distancias dan grandes saltos. El conocimiento que el analista tenga acerca del problema decidirá cuáles de los grupos obtenidos son significativos y cuáles no.

De esta manera, el analista podría establecer un tope para detener el proceso a su conveniencia (esta información puede obtenerse del gráfico del dendrograma). Por ejemplo, podría hacerlo cuando la distancia entre los grupos exceda un valor específico o cuando las distancias sucesivas entre los pasos marquen un repentino salto.

Por otro lado, para el caso del análisis cluster no jerárquico, se puede trazar un gráfico que compare el número de grupos con la relación entre la varianza total de los grupos y la varianza entre los grupos. El punto del gráfico donde se presente un recodo o doblez marcado indicará el número apropiado de grupos. Por lo general, no merecerá la pena aumentar el número de grupos más allá de este punto. Otra posibilidad para decidir el número óptimo de grupos es definir algún tipo de conceptualización intuitiva de la relación teórica de los datos.

Otro problema que puede presentarse en este tipo de análisis es la presencia de grupos unipersonales, es decir, clusters formados por un solo individuo. Son un problema porque podrían ser outliers (valores atípicos) no detectados en el proceso de depuración de nuestra fuente de datos. Si aparece un grupo de un solo miembro, el analista debe estudiar si representa un componente estructural válido en la muestra o si, por el contrario, debiera suprimirse por no ser representativo. Si se suprime del análisis alguna

observación, el investigador deberá ejecutar de nuevo el análisis cluster para las nuevas observaciones válidas y conseguir así definir nuevos grupos.

Métodos gráficos

Para definir el número de conglomerados son muy útiles los gráficos de carámbano o témpanos verticales y el dendograma, disponibles en el programa estadístico SSPS, que es una manera muy ilustrativa, porque muestran a cada paso la formación o agrupación de los objetos que pasan a formar los conglomerados.

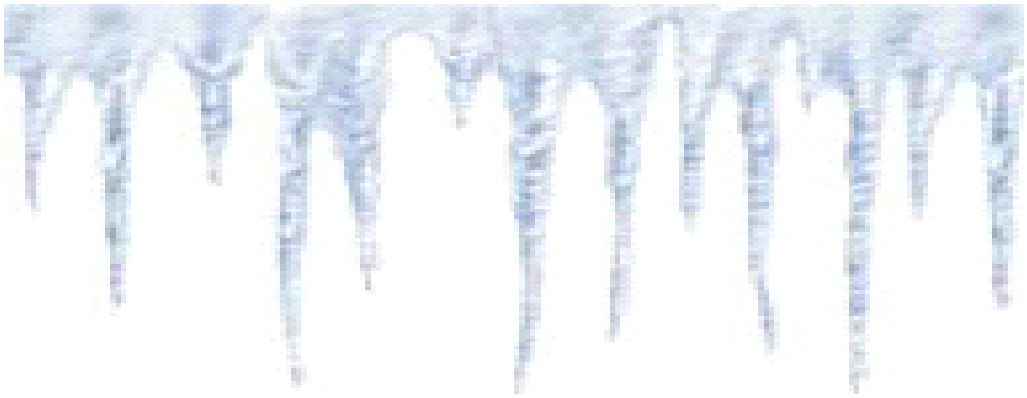
a) Gráfico de Carámbanos o Témpanos verticales (Vertical Icicle Plot)

Es la figura que se utiliza para representar la formación de conglomerados durante su análisis. Su nombre procede de su parecido con una fila de témpanos colgando del alero.

Diagrama de Témpanos Verticales

Número de conglomerado	Casos: Variación porcentual de los Países de Latinoamérica, 2007																	
	7:ECUADOR		5:BOLIVIA		4:CHILE		3:BRASIL		9:PARAGUAY		10:URUGUAY		6:COLOMBIA		8:VENEZUELA		2:ARGENTINA		1:PERU
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X
3	X		X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X
4	X		X	X	X	X	X		X	X	X	X	X		X	X	X	X	X
5	X		X	X	X		X		X	X	X	X	X		X	X	X	X	X
6	X		X	X	X		X		X		X	X	X		X	X	X	X	X
7	X		X	X	X		X		X		X	X	X		X		X	X	X
8	X		X	X	X		X		X		X	X	X		X		X		X
9	X		X		X		X		X		X	X	X		X		X		X

Carámbanos o Témpanos



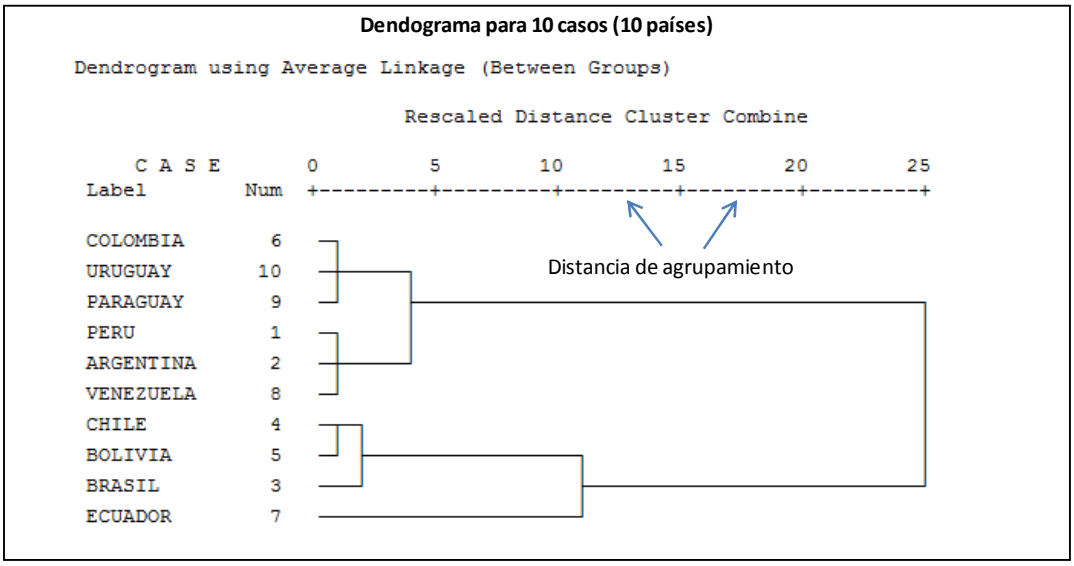
Los casos constituyen las columnas de la visualización y los pasos las filas. Sus columnas corresponden a los objetos que se agrupan y las filas al número de grupos. Esta figura se lee de abajo hacia arriba. Inicialmente todos los casos se consideran como grupos individuales. En el primer caso, se combinan los dos objetos más cercanos.

Los casos que se unen horizontalmente constituyen los conglomerados.

b) Dendrograma o Gráfico de Árbol (Dendogram)

Mediante este gráfico, se sugieren agrupaciones crecientes en número de casos. Se puede leer de izquierda a derecha (para el gráfico siguiente es el indicado) o de abajo hacia arriba. Las líneas verticales representan los grupos unidos. La posición de la línea en la escala indica las distancias en las que se unen los grupos. El dendrograma es un gráfico que muestra el historial de conglomeración en el cual los conglomerados están representados mediante trazos horizontales, las etapas en que se juntan mediante trazos verticales. Este grafico nos ayuda a evaluar la homogeneidad de los conglomerados y nos permite decidir sobre cual es el número óptimo de conglomerados.

Dendograma para 10 casos (10 países)



2.3.7. ANÁLISIS DE LOS RESULTADOS

Como la necesidad es de clasificar una gran cantidad de información en grupos manejables y con significado, el asunto es donde “cortar” el gráfico de manera que se obtenga un número óptimo de grupos. Gran parte de la dificultad para resolver este problema es la ausencia de una hipótesis nula, debido a la falta de una definición consistente de la estructura y contenido de un conglomerado. Una posible hipótesis nula sería: “no estructura” en un conjunto de datos, pero esta está lejos de ser una proposición clara y fácilmente manejable. En cualquier caso, los procedimientos heurísticos son los métodos más utilizados para decidir el nivel de “corte” en un dendograma, mediante la inspección subjetiva de los diferentes niveles del árbol. Además, para interpretar la clasificación obtenida por un Análisis Cluster se requiere, en primer lugar, un conocimiento suficiente del problema analizado. Hay que estar abierto a la posibilidad de que no todos los grupos obtenidos tienen por qué ser significativos

El propósito de esta etapa es, esencialmente, examinar la variación de los conglomerados para asignar etiquetas que describan de un modo veraz su naturaleza. Resulta útil elaborar el perfil de los grupos en términos de las variables utilizadas para el conglomerado. Con respecto al perfilado de los conglomerados o grupos, cabe indicar que no es más que la descripción de las características de cada conglomerado para explicar cómo podrían inferir en dimensiones relevantes. El analista utiliza los datos no incluidos previamente en el procedimiento de aglomeración para perfilar las características de cada conglomerado.

En conclusión, los perfiles se enfocan a describir no a lo que determinan directamente los conglomerados sino (una vez se han determinado los distintos grupos) a sus características propias. Por ello, se hace especial énfasis en las características que definen los grupos y en la capacidad de los miembros de cada conglomerado para predecir una actitud particular del conglomerado en cuestión.

CAPÍTULO III

ANÁLISIS MULTIVARIANTE DE LOS INDICADORES DE DESARROLLO HUMANO DE LOS PAÌSES DE AMÉRICA

3.1. GENERALIDADES

En América, el avance democrático y el desarrollo económico y social han tomado senderos distintos durante las últimas décadas, incluso en largos periodos parecen haberlo hecho en sentidos opuestos. En este contexto la problemática del desarrollo (discutido en los años sesentas en torno a "desarrollo versus subdesarrollo") es una tarea pendiente de soluciones que permitan una evolución sostenida a largo plazo. Para tener un enfoque que puedan permitir analizar la jerarquización del nivel de desarrollo humano en los países de América se requiere de una serie de metodologías estadísticas que permitan la clasificación e identificación de los países integrantes.

El Análisis Multivariante permite estudiar en su conjunto la información disponible para analizar el Índice de Desarrollo Humano, responder a algunas de las preguntas formuladas y confirmar algunas de las conclusiones presentadas en el Informe Sobre Desarrollo Humano 2007 - 2008 del PNUD. Dada la naturaleza cuantitativa de la información el uso del Análisis Factorial se justifica para la sintetización de los indicadores para luego segmentar a los países mediante el análisis cluster. La técnica permite ir formando grupos homogéneos de conglomerados.

3.2. FUENTE DE DATOS

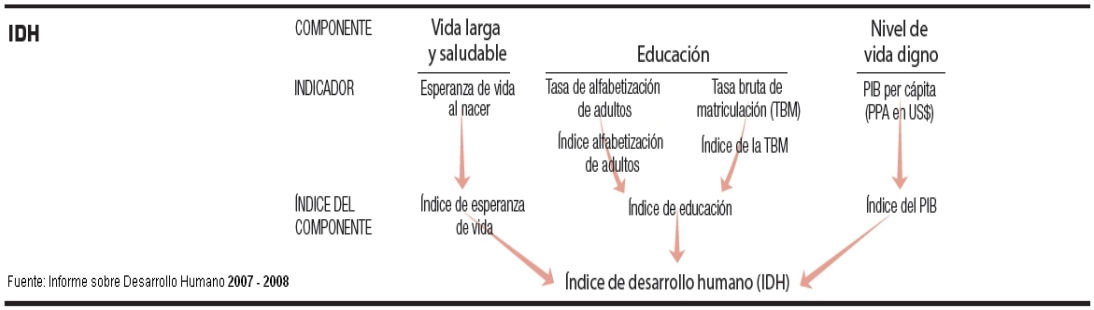
La fuente de datos que se usa en el presente trabajo ha sido obtenida del INFORME SOBRE DESARROLLO HUMANO 2007 – 2008[15], publicado por el Programa de las Naciones Unidas para el Desarrollo (PNUD)[16].

Sin embargo, según consigna la oficina encargada del Informe sobre Desarrollo Humano, esta oficina es principalmente usuaria de estadísticas, no productora. Por ello, se basa en la información que le facilitan organismos internacionales que proveen esta información y que disponen de los recursos y conocimientos necesarios para recopilar y procesar datos internacionales sobre indicadores estadísticos específicos. Las fuentes de todos los datos utilizados en los cuadros de indicadores se citan en una nota breve al final de cada cuadro. Estas notas corresponden a las referencias completas que se presentan en las referencias estadísticas. Para conocer información técnica más detallada acerca de estos indicadores, se pueden hacer en el sitio Web del Informe sobre Desarrollo Humano ubicado en http://hdr.undp.org/statistics/.

3.3. CÁLCULO DEL ÍNDICE DE DESARROLLO HUMANO

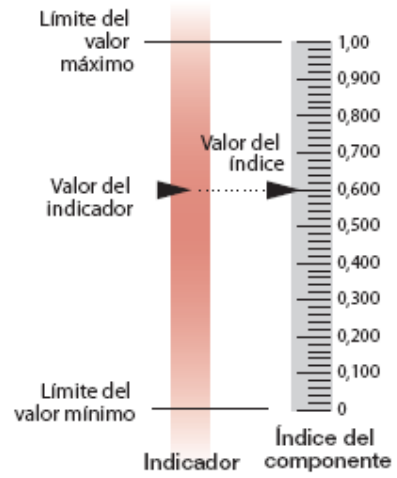
Para calcular el IDH es necesario crear un índice para cada una de sus dimensiones o componentes (los índices de esperanza de vida, educación y PIB), para lo cual se escogen valores mínimos y máximos (valores límite) para cada indicador.

Secuencia del cálculo del IDH



[15] http://hdr.undp.org/en/media/hdr_20072008_sp_complete.pdf
[16] http://www.undp.org/spanish/

Límites de valor del IDH



Fuente: Informe sobre Desarrollo Humano

El desempeño de cada componente se expresa como valor entre 0 y 1, para cuyo efecto se aplica la siguiente fórmula general:

$$\text{Índice del componente} = \frac{\text{valor real - } \mathbf{\text{valor mínimo}}}{\mathbf{\text{valor máximo}} - \mathbf{\text{valor mínimo}}}$$

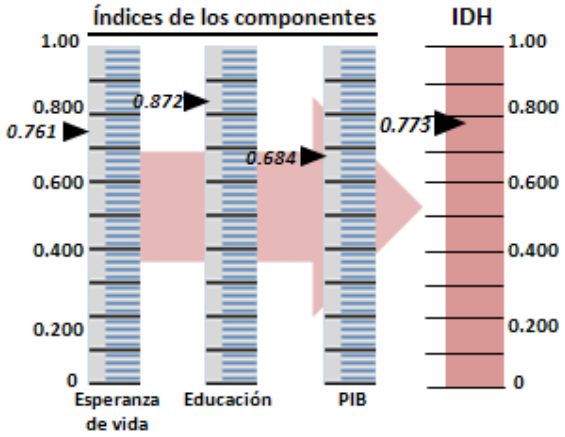
Luego, se calcula el IDH como simple promedio de los índices de los componentes.

Los datos corresponden al Perú:

Cálculo del Índice Desarrollo Humano

Una vez que se han calculado los índices de cada componente, la determinación del IDH es sencilla. Se halla el promedio simple de los índices de los tres componentes.

IDH = 1/3(índice de esperanza de vida) + 1/3(índice de educación) + 1/3(índice del PIB)
= 1/3 (0.761) + 1/3 (0.872) + 1/3(0.684) **= 0.773**



En el subcapítulo 3.4, se muestra el cálculo de los índices para cada componente del IDH de Perú.

En el siguiente cuadro se muestra los valores límites para cada indicador que integran el IDH.

Valores límites para el cálculo del IDH

Indicador	Valor máximo	Valor mínimo
Esperanza de vida al nacer (años)	**85**	**25**
Tasa de alfabetización de adultos (%)*	**100**	**0**
Tasa bruta combinada de matriculación (%)	**100**	**0**
PIB per cápita (PPA en US$)	**40,000**	**100**

Fuente: Informe sobre Desarrollo Humano **2007 - 2008 (PNUD)**

*El valor límite para calcular la alfabetización de adultos implica que la tasa máxima es 100%. En la práctica, para el cálculo del IDH el límite de las tasas de alfabetización superiores a 99% se fija en 99%.

A continuación se realiza una descripción de los indicadores, de acuerdo a la tabla anterior.

3.3.1. Esperanza de vida al nacer

Según PNUD en las Definiciones de los términos estadísticos[17], de define a la Esperanza de Vida al nacer como, número de años que vivirá un recién nacido si los patrones de mortalidad por edades imperantes en el momento de su nacimiento siguieran siendo los mismos a lo largo de toda su vida. Además, según CELADE, la esperanza de vida al nacer se define como el número de años que en promedio esperaría vivir una persona si durante toda su vida estuviera sujeta a las condiciones de mortalidad por edad observadas en el período de estudio.

3.3.2. Tasa de alfabetización de adultos

El Informe de IDH 2007 – 2008, utiliza datos sobre alfabetización de adultos del estudio de abril de 2007 realizado por el Instituto de Estadísticas (UIS) de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO, Instituto de Estadísticas 2007a), el que combina cálculos nacionales directos con cálculos recientes basados en su modelo de proyecciones mundiales de la alfabetización específica por edad creado en

[17] Informe sobre Desarrollo Humano 2007-2008 - PNUD, página 383

2007. Una vez obtenidos niveles superiores de alfabetización, muchos países de ingresos altos dejan de recopilar estadísticas básicas sobre alfabetización. Por lo tanto, para calcular el Índice de Desarrollo Humano de esos países, se aplica una tasa de alfabetización de 99%. Para recopilar datos sobre alfabetización, muchos países calculan la cantidad de personas alfabetizadas sobre la base de datos informados por los propios involucrados. Otros recurren a datos de logros educacionales como una variable substitutiva, pero las mediciones de la asistencia escolar o de la terminación de un curso o grado pueden variar. Los valores para este estudio, se refieren a la tasa de Alfabetización de Adultos en porcentaje de personas mayores a 15 años.

3.3.3. Tasa bruta combinada de matriculación en enseñanza primaria, secundaria y terciaria (Escolaridad)

El UIS (Instituto de Estadísticas de la UNESCO 2007c) produce tasas brutas de matriculación (Tasa bruta combinada de matriculación en enseñanza primaria, secundaria y terciaria) a partir de datos recopilados por los gobiernos nacionales (por lo general, de fuentes administrativas) y datos demográficos de World Population Prospects 1950–2040: The 2004 Revision. Las tasas se calculan dividiendo la cantidad de estudiantes matriculados en los niveles de la enseñaza primaria, secundaria y terciaria por la población total del grupo que tiene la edad oficial correspondiente a esos niveles. El grupo en edad del nivel terciario se fija en las cinco cohortes inmediatamente siguientes al término del ciclo superior de la enseñanza secundaria en todos los países.

Aunque está concebida como variable substitutiva del nivel de instrucción, la tasa bruta combinada de matriculación no refleja los resultados y la calidad de la enseñanza e incluso cuando se usa para capturar el acceso a las oportunidades educativas, puede encubrir diferencias importantes entre los países debido a disparidades en la escala de edades que corresponde a un determinado nivel de educación y en la duración de los programas de enseñanza. Por otra parte, las tasas de repetición de cursos y de deserción escolar también pueden generar distorsiones en los datos. Hay mediciones que captan mejor los resultados de la educación, como los años medios de

escolaridad de una población o la esperanza de escolaridad, y sería ideal que éstas reemplazaran las tasas brutas de matriculación en el Índice de Desarrollo Humano. Sin embargo, ese tipo de datos aún no está disponible de manera sistemática para una cantidad suficiente de países. Según su definición actual, la tasa bruta combinada de matriculación de un determinado país no toma en cuenta a los estudiantes que cursan estudios fuera de su país de origen. Los datos actuales de muchos países pequeños, cuyos habitantes suelen acudir al extranjero a cursar sus estudios terciarios, podrían traducirse en una sub-representación importante del acceso a la educación o del nivel de instrucción de su población y así arrojar un valor de Índice de Desarrollo Humano más bajo.

3.3.4. PBI per cápita (PPA en US$)

Al comparar el nivel de vida entre los países, es necesario convertir las estadísticas económicas a su paridad del poder adquisitivo (PPA) para neutralizar las diferencias en los niveles de precios entre los distintos países. El Banco Mundial (Banco Mundial 2007b) aportó datos sobre el PIB per cápita (PPA en US$) de 168 países para calcular el Índice de Desarrollo Humano, datos que se basan en información sobre precios de las últimas encuestas del Programa de Comparación Internacional (PCI) y en el PIB en moneda local proveniente de datos de las cuentas nacionales. La última ronda de encuestas del programa PCI se realizó entre 1993 y 1996 y abarcó 118 países. Las PPA de estos países se calcularon directamente extrapolando los resultados de referencia más recientes.

3.4. DESCRIPCIÓN Y CÁLCULO DE LOS INDICADORES DEL IDH

Los siguientes indicadores del IDH y variables complementarias serán usados para el presente trabajo.

3.4.1. Índice de Esperanza de Vida

Además de los conceptos descritos anteriormente de este indicador, en el Perú, según el Instituto Nacional de Estadísticas e Informática (INEI), la

Esperanza de Vida al Nacer expresa el promedio de años que espera vivir un recién nacido si las condiciones de mortalidad existentes a la fecha persistieran durante toda su vida. Por otro lado, este indicador resume las condiciones de vida de la población; es decir, el ambiente donde se desarrolla, los niveles de ingreso, el acceso a los servicios de salud, así como los servicios principales que debe disponer la vivienda, etc.

Para la presente aplicación, los cálculos de la esperanza de vida al nacer provienen de World Population Prospects (Perspectivas de Población Mundial) 1950-2050, The 2006 Revisión (ONU, 2007e), fuente oficial de las proyecciones y cálculos demográficos de la ONU elaborados cada dos años por la División de Población del Departamento de Asuntos Económicos y Sociales de las Naciones Unidas a partir de datos de los registros demográficos, censos de población y encuestas de los países.

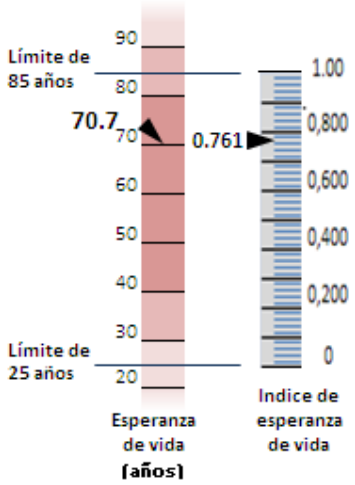
A continuación mostramos el cálculo de este indicador con la información de Perú a modo de un ejemplo:

Los datos corresponden al país de Perú:

Cálculo del Índice de esperanza de vida

Este indicador mide los logros de un país en cuanto a esperanza de vida al nacer. Para Perú, cuya esperanza de vida era de 70.7 años en el 2005, el índice de esperanza de vida es de 0.761.

$$\text{Índice de esperanza de vida} = \frac{70.7 - 25}{85 - 25} = 0.761$$



3.4.2. Índice de Educación

El índice de educación mide el progreso relativo de un país en materia de alfabetización de adultos y matriculación bruta combinada en educación primaria, secundaria y terciaria. En primer lugar, se calcula el índice de alfabetización de adultos y el índice de la tasa bruta combinada de matriculación. Luego, se combinan ambos índices para crear el índice de educación, en el que se otorga una ponderación de dos terceras partes a la

alfabetización de adultos y de una tercera parte a la tasa bruta combinada de matriculación.

A modo de ejemplo, mostramos el cálculo de este indicador para Perú:

Cálculo del Índice de educación

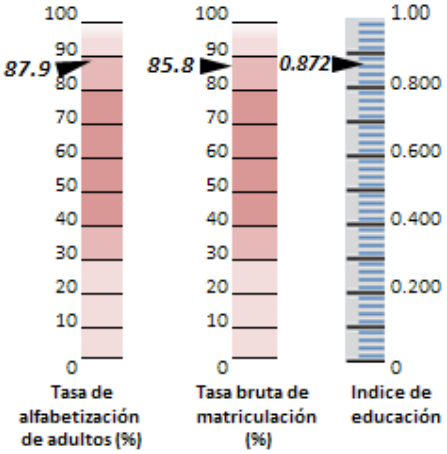
Para Perú, cuya tasa de alfabetización de adultos en 2005 era de 87.9% y la tasa bruta combinada de matriculación era de 85.8% en el mismo año, el índice de educación es de 0.872

Índice de alfabetización de adultos = $\frac{87.9 - 0}{100 - 0}$ = **0.879**

Índice bruto de matriculación = $\frac{85.8 - 0}{100 - 0}$ = **0.858**

Índice de educación = 2/3(índice de alfabetización de adultos) + 1/3(índice de matriculación bruta)

= 2/3 (0.879) + 1/3 (0.858) = **0.872**



3.4.3. Índice del PBI

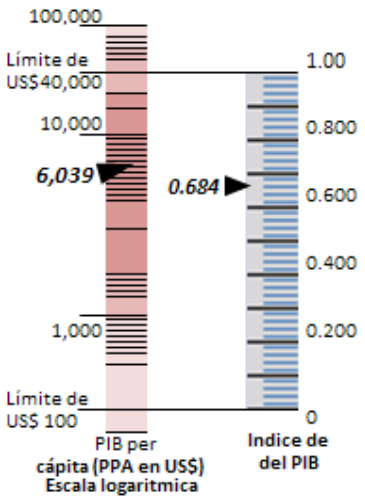
El índice del PBI se calcula utilizando el PBI per cápita ajustado (PPA[18] en US$). En el IDH, los ingresos actúan como sustitutos de todos los demás componentes del desarrollo humano que no se reflejan en una vida larga y saludable ni en la educación adquirida. Los ingresos se ajustan puesto que lograr un nivel respetable de desarrollo humano no requiere ingresos ilimitados.

El cálculo es de la siguiente manera:

Cálculo del Índice del PBI

Para Perú, cuya cuyo PIB per cápita fue de $ 6,039 (PPA US$) en 2005, el índice del PIB es de 0.684

Índice del PBI = $\frac{\log(6{,}039) - \log(100)}{\log(40{,}000) - \log(100)}$ = **0.684**



[18] PPA: Paridad de Poder Adquisitivo. Sirve de base para la elaboración del PBI Per Cápita de cada país, para efectos de comparación y otros usos estadísticos – económicos. Los cálculos de la PPA son útiles para comparar el nivel de vida real en diferentes países.

3.4.4. Tasa de mortalidad infantil

Se define como la probabilidad de morir entre el parto y al cumplir exactamente un año de edad, expresada por cada 1.000 nacidos vivos. Es decir, representa la frecuencia de muertes de niños menores de un año de edad por cada mil niños nacidos vivos.

3.4.5. Tasa de mortalidad de niños menores de cinco años

Es la probabilidad de morir entre el parto y exactamente los cinco años de edad, expresada por cada 1.000 nacidos vivos, en un determinado año. Además, esta tasa puede usarse para reflejar tanto las tasas de mortalidad infantil como de la niñez. Esto tiene ciertas ventajas. El uso de solo la tasa de mortalidad infantil puede distraer la atención de una elevada tasa de mortalidad entre niños mayores. Los problemas de malnutrición, en particular, pueden perder la atención debida como factor causal, especialmente en aquellos países donde el niño es más vulnerable a la malnutrición en su segundo año de vida.

3.4.6. Gasto en salud, Sector Público

Es el porcentaje del PBI que invierte los gobiernos compuesto por: gasto corriente y de capital con cargo al presupuesto gubernamental (gobiernos centrales y locales), préstamos y subvenciones externas (incluidas donaciones de organismos internacionales y organizaciones no gubernamentales) y fondos de seguros de salud, de carácter social (u obligatorio). Junto con el gasto privado en salud, constituye el gasto en salud total.

3.4.7. Gasto en salud, Per Cápita

Representa el gasto en salud per cápita (PPA en US$), compuesto por la suma del gasto público y privado (PPA en US$), dividido por la población correspondiente. El gasto en salud comprende prestación de servicios sanitarios (preventivos y curativos), actividades de planificación familiar,

actividades de nutrición y de ayuda en caso de emergencia destinada a la salud, pero no incluye suministro de agua y saneamiento.

Las variables anteriormente descritas se resumen en la siguiente tabla:

INDICADOR	DESCRIPCIÓN	CATEGORIA
Índice de esperanza de vida.	Este indicador mide los logros de un país en cuanto a esperanza de vida al nacer. La esperanza de vida es el promedio de años que se espera que viva un recién nacido de acuerdo con la probabilidad de sobrevivencia prevaleciente en el momento del nacimiento.	Población
Índice de la educación.	Progreso relativo de un país en materia de alfabetización de adultos y matriculación bruta combinada en educación primaria, secundaria y terciaria.	Educación
Índice del PBI.	El índice del PBI se calcula utilizando el PBI per cápita ajustado (PPA en US$). El PBI es el valor monetario de todos los bienes y servicios finales que se producen en una economía en un período de tiempo determinado.	Economía
Tasa de mortalidad infantil.	Probabilidad de morir entre el parto y al cumplir exactamente un año de edad, expresada por cada 1.000 nacidos vivos.	Población
Tasa de mortalidad infantil menores de 5 años.	Probabilidad de morir entre el parto y exactamente los cinco años de edad, expresada por cada 1.000 nacidos vivos, en un determinado año.	Población
Gasto en salud pública, % del PBI.	Es el porcentaje del PBI que invierte los gobiernos. Gasto corriente y de capital con cargo al presupuesto gubernamental, préstamos y subvenciones externas y fondos de seguros de salud, de carácter social.	Salud
Gasto en salud, per cápita en dólares.	Gasto en salud per cápita (PPA en US$). Suma del gasto público y privado (PPA en US$), dividida por la población de medio año.	Salud

Fuente: Informe sobre el Desarrollo Humano 2007 - 2008, PNUD

3.5. PROCESAMIENTO DE LOS DATOS

Para el procesamiento de los datos, se estructuró una matriz para el ingreso de la información; se realizó una verificación de la calidad de la digitación al 100%, mediante doble digitación, para asegurar error “cero” en el

ingreso de la información. Se utilizó los programas estadísticos SPSS versión 15 y Minitab 14.

CUADRO N° 1

Índice de Desarrollo Humanos y sus componentes de los países americanos

INFORME 2007 - 2008

ZONA	PAIS	RANKING	CLASIFICACION	IDH	Componentes del IDH				Índice de los componentes del IDH y variables complementarias						
					ESPERANZA DE VIDA (AÑOS)	TASA DE ALFABETISMO ADULTO	ESCOLARIDAD	PIB PER CAP ($)	INDICE DE ESPERANZA DE VIDA	INDICE DE EDUCACION	INDICE DEL PBI	Tasa de mortalidad infantil	Tasa de mortalidad infantil < de 5 años	Gasto en salud publica, %del PBI	Gasto en salud, per capita ($)
América del norte	CANADA	4	A	**0.961**	80.3	99	99.2	33375	0.921	0.991	0.97	5	6	6.8	3173
	ESTADOS UNIDOS	12	A	**0.951**	77.9	99	93.3	41890	0.881	0.971	1	6	7	6.9	6096
	MEXICO	52	A	**0.829**	75.6	91.6	75.6	10751	0.843	0.863	0.781	22	27	3	655
América central	ANT. Y BARBUDA	57	A	**0.815**	73.9	85.8	76	12500	0.815	0.824	0.806	11	12	3.4	516
	BAHAMAS	49	A	**0.845**	72.3	95.8	70.8	18380	0.789	0.875	0.87	13	15	3.4	1349
	BARBADOS	31	A	**0.892**	76.6	99.7	88.9	17297	0.861	0.956	0.86	11	12	4.5	1151
	BELICE	80	M	**0.778**	75.9	75.1	81.8	7109	0.849	0.773	0.712	15	17	2.7	339
	COSTA RICA	48	A	**0.846**	78.5	94.9	73	10180	0.891	0.876	0.772	11	12	5.1	592
	CUBA	51	A	**0.838**	77.7	99.8	87.6	6000	0.879	0.952	0.683	6	7	5.5	229
	DOMINICA	71	M	**0.798**	75.6	88	81	6393	0.844	0.857	0.694	13	15	4.2	309
	EL SALVADOR	103	M	**0.735**	71.3	80.6	70.4	5255	0.772	0.772	0.661	23	27	3.5	375
	GRANADA	82	M	**0.777**	68.2	96	73.1	7843	0.72	0.884	0.728	17	21	5	480
	GUATEMALA	118	M	**0.689**	69.7	69.1	67.3	4568	0.746	0.685	0.638	32	43	2.3	256
	HAITI	146	M	**0.529**	59.5	54.8	53	1663	0.575	0.542	0.469	84	120	2.9	82
	HONDURAS	115	M	**0.7**	69.4	80	71.2	3430	0.739	0.771	0.59	31	40	4	197
	JAMAICA	101	M	**0.736**	72.2	79.9	77.9	4291	0.787	0.792	0.627	17	20	2.8	223
	NICARAGUA	110	M	**0.71**	71.9	76.7	70.6	3674	0.782	0.747	0.601	30	37	3.9	231
	PANAMA	62	A	**0.812**	75.1	91.9	79.5	7605	0.836	0.878	0.723	19	24	5.2	632
	REP. DOMINICANA	79	M	**0.779**	71.5	87	74.1	8217	0.776	0.827	0.736	26	31	1.9	377
	SAN CRISTOBAL	54	A	**0.821**	70	97.8	73.1	13307	0.75	0.896	0.816	18	20	3.3	710
	SANTA LUCIA	72	M	**0.795**	73.1	94.8	74.8	6707	0.802	0.881	0.702	12	14	3.3	302
	SAN VICENTE	93	M	**0.761**	71.1	88.1	68.9	6568	0.768	0.817	0.698	17	20	3.9	418
América del sur	ARGENTINA	38	A	**0.869**	74.8	97.2	89.7	14280	0.831	0.947	0.828	15	18	4.3	1274
	BOLIVIA	117	M	**0.695**	64.7	86.7	86	2819	0.662	0.865	0.557	147	65	4.1	186
	BRASIL	70	A	**0.8**	71.7	88.6	87.5	8402	0.779	0.883	0.74	31	33	4.8	1520
	CHILE	40	A	**0.867**	78.3	95.7	82.9	12027	0.889	0.914	0.799	8	10	2.9	720
	COLOMBIA	75	M	**0.791**	72.3	92.8	75.1	7304	0.788	0.869	0.716	17	21	6.7	570
	ECUADOR	89	M	**0.772**	74.7	91	75	4341	0.828	0.858	0.629	22	25	2.2	261
	GUYANA	97	M	**0.75**	65.2	99	85	4508	0.67	0.943	0.636	47	63	4.4	329
	PARAGUAY	95	M	**0.755**	71.3	93.5	69.1	4642	0.771	0.853	0.641	20	23	2.6	327
	PERU	87	M	**0.773**	70.7	87.9	85.8	6039	0.761	0.872	0.684	23	27	1.9	235
	SURIMAN	85	M	**0.774**	69.6	89.6	77.1	7722	0.743	0.854	0.725	30	39	3.6	376
	URUGUAY	46	A	**0.852**	75.9	96.8	88.9	9962	0.848	0.942	0.768	14	15	3.6	784
	VENEZUELA	74	M	**0.792**	73.2	93	75.5	6632	0.804	0.872	0.7	18	21	2	285

Fuente: Informe sobre Desarrollo Humano 2007 - 2008 (PNUD) Elaboración propia

Clasificación: A= Alto M= Medio B=Bajo

Se seleccionaron a los 34 países del continente americano, que llamaremos Países Americanos para este estudio (incluye El Caribe), de un total de 177 estados miembros de las Naciones Unidas para los cuales pudo calcularse el IDH. Los países aparecen por orden alfabético (en su zona

americana) según el IDH. La mayoría de los datos que se presentan en los cuadros se refiere al año 2005 y son aquellos que estaban disponibles en la oficina encargada del Informe sobre Desarrollo Humano el 1 de julio de 2007, a menos que se indique otra cosa.

CUADRO Nº 02: ESTADISTICOS DESCRIPTIVOS DEL IDH Y SUS COMPONENTES DE LOS PAISES AMERICANOS

Estadísticos descriptivos

	N	Mínimo	Máximo	Media
IDH (índice)	34	.53	.96	.79
ESPERANZA (años)	34	59.50	80.30	72.64
ALFABETISMO (%)	34	54.80	99.80	89.33
ESCOLARIDAD (%)	34	53.00	99.20	78.20
PBI per cáp. (PPA en $)	34	1,663	41,890	9,579

Fuente: Informe sobre Desarrollo Humano 2007 - 2008 (PNUD)

El menor IDH de los 34 países americanos pertenece a Haití (0.53) y Canadá presenta el IDH más alto (0.96). En promedio el IDH de los países americanos (0.79), se acerca a la clasificación Alto (que es desde 0.80). En general, todos los valores mínimos del Cuadro N° 02 pertenecen al país de Haití, los valores máximos pertenecen a Canadá y EE.UU.

CUADRO Nº 03: ESTADISTICOS DESCRIPTIVOS DE LOS ÍNDICES DE LOS COMPONENTES DEL Y VARIABLES COMPLEMENTARIAS DE LOS PAISES AMERICANOS

Estadísticos descriptivos

	N	Mínimo	Máximo	Media
Índice de la Esperanza de Vida	34	.575	.921	.79412
Índice de educación	34	.542	.991	.85594
Índice del PBI	34	.469	1.000	.72235
Gasto público en salud en % del PBI	34	1.9	6.9	3.841
Gasto en salud per cápita en PPA en $	34	82	6096	751.74
Tasa mortalidad infantil < de 1 año	34	5	147	24.44
Tasa mortalidad infantil < de 5 años	34	6	120	26.68

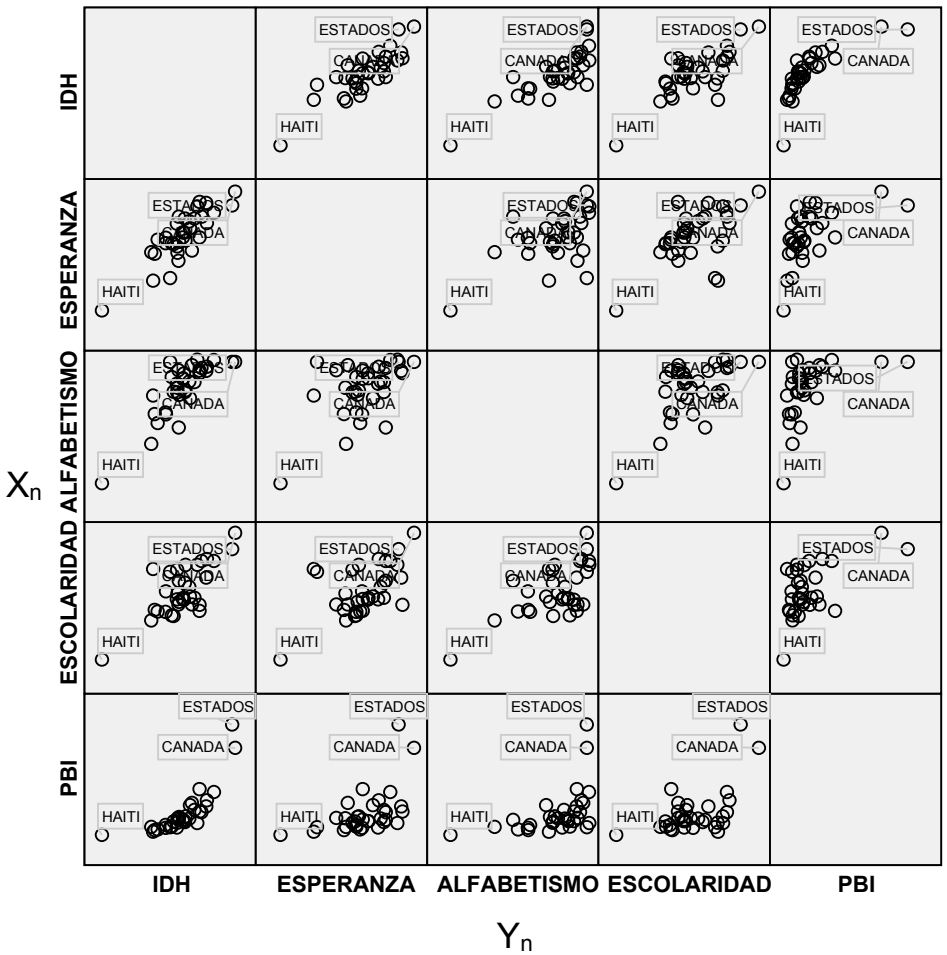
Fuente: Informe sobre Desarrollo Humano 2007 - 2008 (PNUD)

Al igual que la descripción anterior, Haití presenta los valores más bajos en lo que se refiere al Índice de Esperanza de Vida, de educación y del PBI, además es seguido por Bolivia. Cabe señalar, que los países de centro América en general presentan los menores índices de educación. Perú y

República Dominicana presentan el menor porcentaje del PBI que invierte en salud pública (1.9% del PBI).

La representación gráfica que permite comprobar la existencia de relación lineal entre dos variables es el diagrama de dispersión. Un gráfico de dispersión para un análisis de varias variables que ofrece mucha información es el que se representa en una estructura matricial de todos los pares posibles de variables, como se observa en el gráfico N°01, la mayoría de los pares presentan una relación lineal positiva; es decir, a medida que aumenta el valor de la variable X_n aumenta también el valor de la variable Y_n.

GRÁFICO Nº 01: DISPERSIÓN DE LOS DATOS DEL IDH Y SUS COMPONENTES DE LOS PAISES AMERICANOS



Además, en el gráfico anterior, cabe resaltar la detección de posibles atípicos, el país de Haití se distancia de demás países, igual pero

opuestamente, Estados Unidos y Canadá se alejan del agrupamiento de los países, este comportamiento se fundamenta en el desarrollo integral de los países del norte de América (EE.UU. y Canadá).

3.6. SINTETIZACIÓN DE LOS INDICADORES DE DESARROLLO HUMANO MEDIANTE EL ANÁLISIS FACTORIAL

La esencia de los métodos factoriales es que consienten una pérdida de información para obtener una ganancia en interpretación, en el presente trabajo se aplica para reducir los conceptos de siete indicadores a menor cantidad de los mismos, que permitirá un mejor análisis y caracterización de los Países Americanos de acuerdo a su IDH. Para esto, se estudia la estructura de una tabla de datos de grandes dimensiones y luego suministran una representación simplificada de la misma, pero más significativa, sencilla y fácil de interpretar. Según Pérez (2004), el análisis factorial tiene como objeto simplificar las múltiples y complejas relaciones que puedan existir entre un conjunto de variables observadas X_1, X_2….X_k. Para ello trata de encontrar dimensiones comunes o factores que ligan a las aparentemente no relacionadas variables[19].

Cabe mencionar, que en el análisis factorial se presupone la existencia de ciertas variables no medidas y de interés que, latentes en la tabla de datos, permanecen a la espera de ser halladas. Esta presunción de existencia de variables subyacentes es la condición clave del Análisis Factorial. Se trata de un método estadístico multivariante distinto del Análisis de Componentes Principales aunque con soporte matemático parecido (Pérez, 2004).

Los supuestos básicos que garantizan el adecuado desarrollo del análisis factorial, para este trabajo, cabe resumirlos en dos características:

1).- Tamaño de muestra adecuado. El análisis factorial exige que el tamaño muestral sea elevado, las muestras pequeñas están predispuestas a estimaciones de los coeficientes de correlación infiables. Según Comrey (1973) propone una escala muestral guía que va desde los 50 casos (que considera un tamaño muestral "muy pobre") hasta 1000 (que representa un tamaño

[19] PÉREZ LÓPEZ, CÉSAR. Técnicas de Análisis Multivariante de Datos, Pearson Educación (2004).

muestral “excelente”). Sin embargo, como sucede en otros procedimientos analíticos, lo más preciso es considerar el tamaño muestral en relación con el número de variables a analizar. Tabachnick y Fidell (1989) proponen, como regla, que exista al menos 5 casos por variable. Para la presente aplicación, tenemos que las variables (7 variables) en análisis cuentan con 34 observaciones, por tanto tenemos 7 x 34 = 238 casos, que se considera adecuado para la aplicación del análisis factorial.

2).- Correlaciones entre las variables. A diferencia de otros análisis multivariantes donde la multicolinealidad es un problema, en el análisis factorial se requiere de la existencia de correlación entre las variables. Si las variables están apenas correlacionadas, no tiene sentido la búsqueda de estructuras latentes (llámense factores comunes), que agrupan a variables observadas (o indicadores), correlacionados entre sí, que expresan una misma dimensión del concepto que se mida. Para este trabajo, al analizar la matriz de correlaciones (ver Anexo A11) observamos que existen correlaciones moderadamente altas, entre los indicadores de desarrollo (variables) en su gran mayoría, siendo el mayor para el par de indicadores: *Tasa de mortalidad infantil menores de 5 años (por cada 1000 nacidos vivos)* con el *Índice de Esperanza de Vida* registrando el valor de valor -0.863, que indica una relación negativa u opuesta; es decir, el aumento de la *Tasa de mortalidad infantil menores de 5 años (por cada 1000 nacidos vivos)* le sigue una disminución en el *Índice de Esperanza de Vida* (Anexo A12). El determinante tiende a cero (0.003), esto significa que si es posible realizar el análisis factorial.

Los valores obtenidos en la matriz anti-imagen en términos absolutos son pequeños, observamos también que existen valores altos para la medida de adecuación muestral de cada variable la cuales se encuentran en la diagonal de esta matriz (Anexo A12), según el KMO se encuentran dentro de la categoría de aceptables y buenos.

CUADRO Nº 04: KMO

Medida de adecuación muestral de Kaiser-Meyer-Olkin.	.765

Todo lo anteriormente expuesto nos lleva a concluir que el análisis factorial que se va a aplicar en este trabajo resulta a priori pertinente y nos puede dar resultados satisfactorios.

En el siguiente cuadro de las comunalidades (Cuadro Nº 05), se puede ver que el 84.6% de la variabilidad del Índice de esperanza de vida es compartida con el resto de variables a través de los factores comunes. La comunalidad es la parte de variabilidad de cada variable explicada por lo factores. Antes de la extracción de los factores la comunalidad de cada variable es la unidad, e interesa que después de la extracción siga siendo alta. La variable Gasto en Salud per-cápita en $, presenta la más alta comunalidad (0.943) lo que indica que tiene más en común con las otras variables incluidas en el análisis.

CUADRO Nº 05: COMUNALIDADES

	Inicial	Extracción
Índice de la Esperanza de Vida	1.000	.846
Índice de educación	1.000	.891
Índice del PBI	1.000	.872
Gasto público en salud en % del PBI	1.000	.772
Gasto en salud per cápita en PPA en $	1.000	.943
Tasa mortalidad infantil < de 1 año	1.000	.852
Tasa mortalidad infantil < de 5 años	1.000	.933

Método de extracción: Análisis de Componentes principales.

Por otro lado se observa que la variable de Inversión sobre Gasto en salud pública en % del PBI, comparte la menor variabilidad 77.2% con el resto de las variables a través de los factores comunes, aún siendo el menor de las variables en análisis, este valor es alto.

Usaremos el Método de Componentes Principales para realizar la estimación de los factores. Este método tiene como objetivo resumir la información y así permitir reducir las dimensiones. Además, realizaremos rotación Varimax, con el objetivo de simplificar la estructura factorial y facilitar la interpretación de la matriz factorial, forzando a las variables a definirse más en una dimensión latente, con preferencia a otras.

Como se analiza la estructura de la matriz de correlación se emplea el Método de Kaiser que nos indica el número de factores con autovalores

mayores que 1. En nuestro análisis solo los dos primeros factores tienen valores mayores a la unidad, indicando que estos factores son capaces de explicar el 79.415% de la variabilidad total.

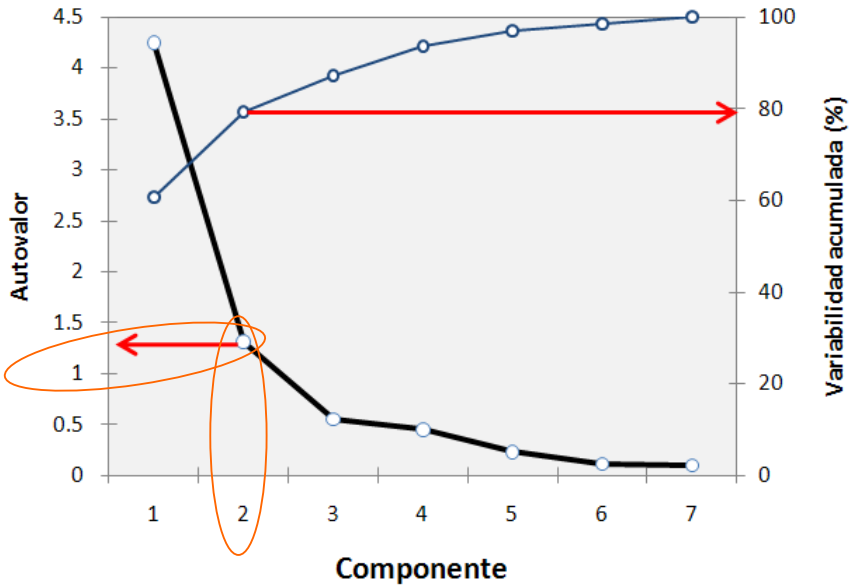
CUADRO Nº 06: VARIANZA TOTAL EXPLICADA.

Componente / Factor	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	4.248	60.686	60.686	4.248	60.686	60.686
2	1.311	18.729	79.415	1.311	18.729	79.415
3	.549	7.842	87.258	.549	7.842	87.258
4	.451	6.437	93.695			
5	.233	3.330	97.025			
6	.111	1.592	98.617			
7	.097	1.383	100.000			

Método de extracción: Análisis de Componentes principales.

Según Catell (1966), el gráfico de sedimentación o gráfico del codo muestra un punto de inflexión pronunciada en la curva a partir del tercer factor, sin embargo solo hasta el segundo factor registra un autovalor mayor a la unidad (el tercer factor solo aporta 7.842%, a diferencia de los dos factores primeros que aportan mayor varianza), lo cual según el cuadro Nº 6 explicaría el 79.415% (como anteriormente se indicó), registrando una cantidad moderada para la expansión de la varianza explicada, lo que nos sugiere elegir 2 factores, esta elección subjetiva depende además del fenómeno que se analiza.

GRÁFICO Nº 02: GRÁFICO DE SEDIMENTACIÓN



Finalmente, tenemos los resultados de la matriz de pesos factoriales, rotada con el método varimax, (rotado con el objetivo de mejorar la observación de agrupación de las variables), esta matriz mide la correlación entre cada variable y el factor. Como podemos observar los pesos factoriales en su gran mayoría son altas, con valores muy próximos a 1, se han eliminado las pesos con valores inferiores a 0.30 para realizar una mejor visualización e interpretar mejor los resultados.

CUADRO Nº 07: MATRIZ DE LAS CARGAS FACTORIALES (ROTADO)

	Componente/Factor	
	1	2
Índice de la Esperanza de Vida	.874	
Tasa mortalidad infantil < de 1 año	-.891	
Tasa mortalidad infantil < de 5 años	-.937	
Gasto público en salud en % del PBI		.858
Gasto en salud per cápita en PPA en $		.859
Índice del PBI	.648	.647
Índice de educación	.558	.580

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

Ahora podemos definir los factores que sintetizan las variables analizadas:

El primer factor, está conformado por las siguientes variables:

- Índice de la Esperanza de Vida.

- Tasa de mortalidad infantil (por cada 1000 nacidos vivos).
- Tasa de mortalidad infantil menores de 5 años (por cada 1000 nacidos vivos).

Estos indicadores se encuentran representados adecuadamente en este factor, que resume el comportamiento vital de las personas de los países de América, a este factor lo denominamos como: **Expectativas de Vida**.

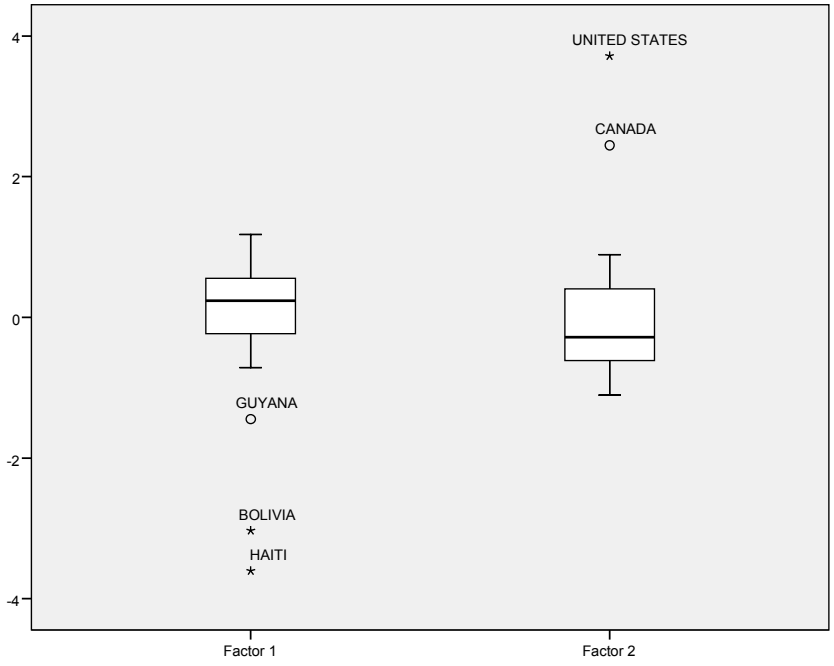
El segundo factor, está conformado por las siguientes variables:

- Gasto en salud per cápita en PPA en $.
- Gasto público en salud en % del PBI.
- Índice del PBI.
- Índice de la educación

El conjunto de estas variables permiten describir el valor monetario de una economía y el gasto y desarrollo de los servicios básicos (salud y educación en los sectores público y privado), a este factor lo llamaremos: **Calidad de Vida.**

Cabe indicar que el primer factor es bipolar, ya que muestra dos cargas negativas. Esto se debe a que existen países con registros altos en las tasas de mortalidad, y las cargas positivas indicaran hay países que tienen una mínima dependencia de estas tasas de mortalidad.

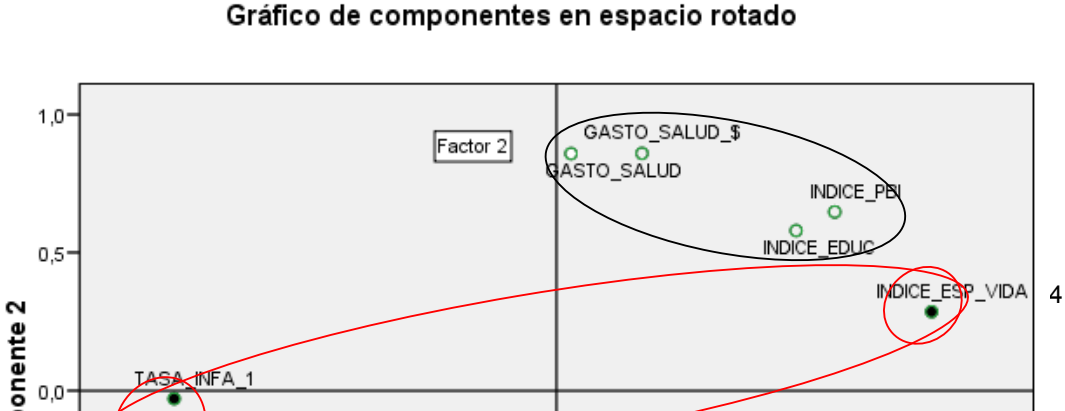
GRÁFICO Nº 03: DIAGRAMA DE CAJAS DE LOS FACTORES (ROTADOS)



El gráfico de cajas nos indica que no hay alta dispersión de los datos, el factor 1 presenta mejor indicio de simetría de los datos (cargas factoriales) en comparación con el factor 2 (tiene la mediana más alejada del centro de la caja). En el factor 1 se observa que Guyana se aleja del extremo del bigote inferior, pero más aún Bolivia y Haití, esto porque son países con expectativas de vida en promedio diferente a los demás. Por otro lado, en el factor 2 (Calidad de Vida), Estados Unidos y Canadá, se alejan del grupo de países, esto dos países se caracterizan porque tienen mejor calidad de vida.

También por medio de los gráficos podemos deducir el sentido y significado de los factores. En el gráfico Nº 4 observamos las coordenadas de cada variable respecto a los factores, que son valores entre -1 y 1 (Cuadro N° 7), siendo mayor la contribución cuanto mayor sea este valor, lo ideal es que no estén centrados en torno al punto (0,0) sino próximos al borde del gráfico. Es así, que podemos apreciar que las tasas de mortalidad infantil y el índice de esperanza de vida, se encuentran correlacionadas negativamente (contrapuestas), esto es, como anteriormente mencionamos, por la bipolaridad de estos índices (dos cargas negativas en las tasas de mortalidad). Esto supone, que mientras menores tasas de mortalidad infantil, aumentará el índice de esperanza de vida. Debemos tener en cuenta que la correlación negativa conlleva que un aumento en una variable va acompañado de una disminución en la otra y viceversa. En general, las variables de mortalidad infantil están correlacionadas negativamente con los dos componentes o factores (con menor fuerza con el componente 2, Calidad de Vida) y todas las demás están correlacionadas positivamente con el componente 1 llamado Expectativas de Vida (en menor fuerza las variables de salud).

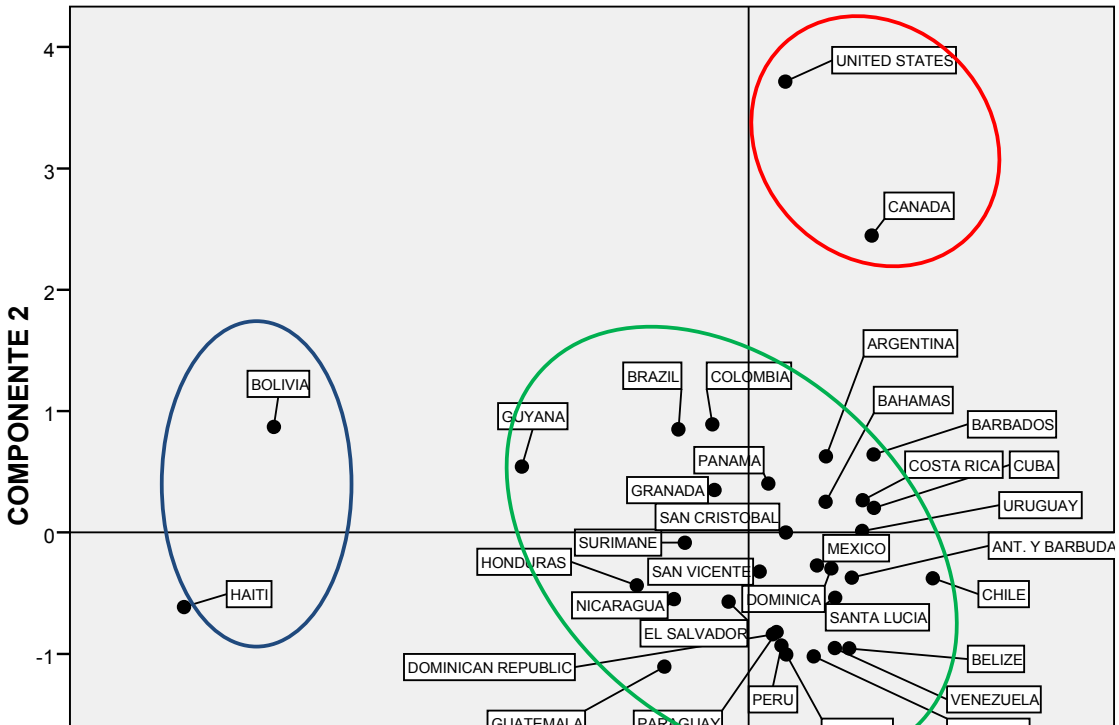
GRÁFICO Nº 04: MAPA FACTORIAL DE LOS INDICADORES DE DESARROLLO HUMANO DE LOS PAÍSES DE AMÉRICA



Una vez concluida con la asignación de las “etiquetas” a las dimensiones latentes (factores hallados), proseguimos con el cálculo de las puntuaciones factoriales, también llamado pesos o cargas factoriales (ver Anexo). Las puntuaciones factoriales se definen como medidas compuestas de cada factor común o componente principal, informan de la posición de cada caso concreto (en nuestro caso países) en cada factor. Destaca su utilidad en la detección de atípicos, en la búsqueda de conglomerados y, en general, para comprender la estructura de los datos. En la presente aplicación, observamos el comportamiento de los pesos factoriales (con procedimiento de regresión) para cada individuo (34 países), en el siguiente plano con los dos factores o componentes.

GRÁFICO Nº 05: PLANO ILUSTRATIVO DE LAS PUNTUACIONES FACTORIALES

Rotación Varimax



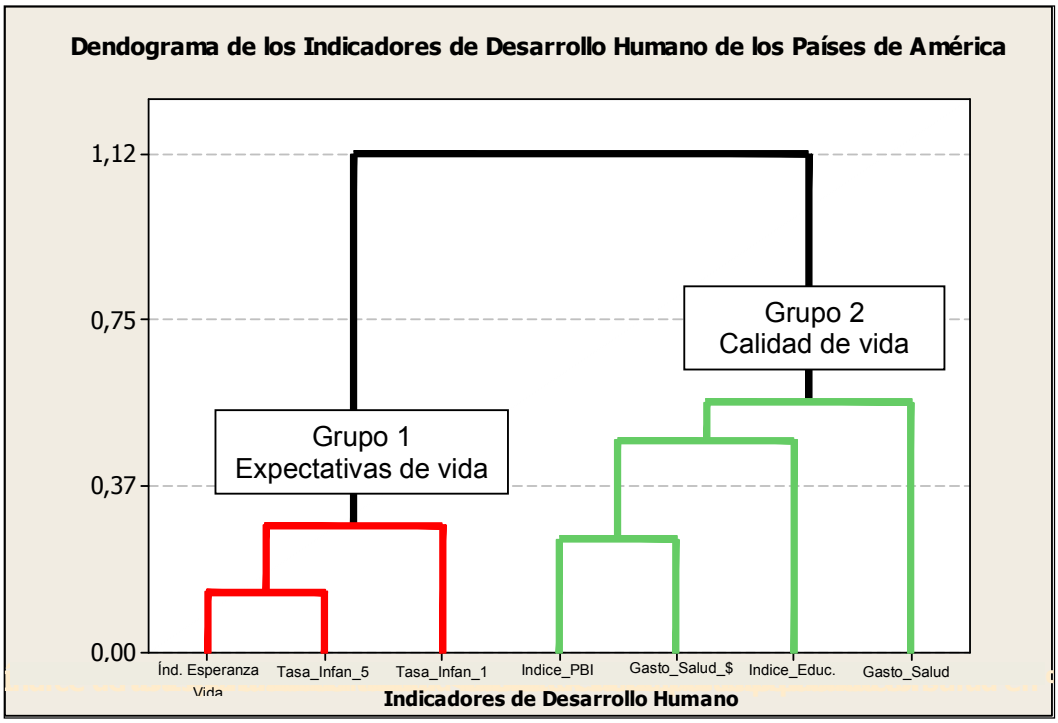
Según D' Ancona (2002), como las puntuaciones factoriales se hallan estandarizadas, se considera atípico al individuo en análisis que supere el valor 2.0, indistintamente de su signo (positivo o negativo), incluso el valor de referencia se eleva a 2.5 o 3.0, dependiendo del tamaño de la muestra. Haití, Bolivia, Estados Unidos y Canadá se identifican como valores atípicos; los dos primeros países se identifican con mayor fuerza con el componente 1; Expectativas de vida (en forma opuesta por los valores negativos), esto indica que Bolivia y Haití se encuentran a más de dos unidades de deviación típica por debajo de la media, esto porque estos países tienen una alta tasa de mortalidad infantil (contrario a lo que se requiere para tener una mejor expectativa de vida). Los dos siguientes países se relacionan más con el componente 2; Calidad de vida, en forma positiva, lo que indica que la calidad de vida en estas naciones es mayor a los demás (presentan más de dos unidades de desviación típica por encima de la media). Además, en el mismo plano, se observa la formación de conglomerados (3 grupos), que a continuación confirmaremos mediante el Análisis Clúster.

3.7. SEGMENTACIÓN DE LOS PAÍSES AMERICANOS MEDIANTE ANÁLISIS CLUSTER

La clasificación tiene como objetivo buscar una estructura "natural" entre las observaciones o variables, preferentemente basadas en un perfil multivariante. La técnica más usada para esta tarea es el Análisis Cluster. Sin embargo, por la característica de este análisis de ser descriptivo, ateórico y no inferencial (Joseph F. Hair, 1999), usaremos de una manera exploratoria, en particular para la comprobación de los grupos de indicadores (factores) y de países hallados cuando se aplicó el Análisis Factorial.

Para confirmar si la agrupación de variables (usadas para el presente análisis) que definieron los factores (2 factores) son adecuados, aplicamos el Análisis Cluster utilizando el algoritmo de división el cual asume que en un primer paso todos los datos conforman un solo cluster, luego para medir la proximidad aplicamos el método de Coeficiente de Correlación Absoluto, debido a que se clasifica variables (absoluto, porque los valores cercano a -1 pueden implicar fuerte semejanza entre las variables). El método de selección es el jerárquico con enlace Ward. El resultado al aplicar el Análisis Cluster hacen suponer la confirmación de los resultados del Análisis Factorial, al mostrar, en el dendograma, similar número de grupos o factores al que muestra en el Análisis Factorial, 2 factores o 2 grupos para el Análisis Cluster, por tanto el nombre general para los grupos son similares al Factorial. El resultado de este procedimiento se representa por el dendograma siguiente:

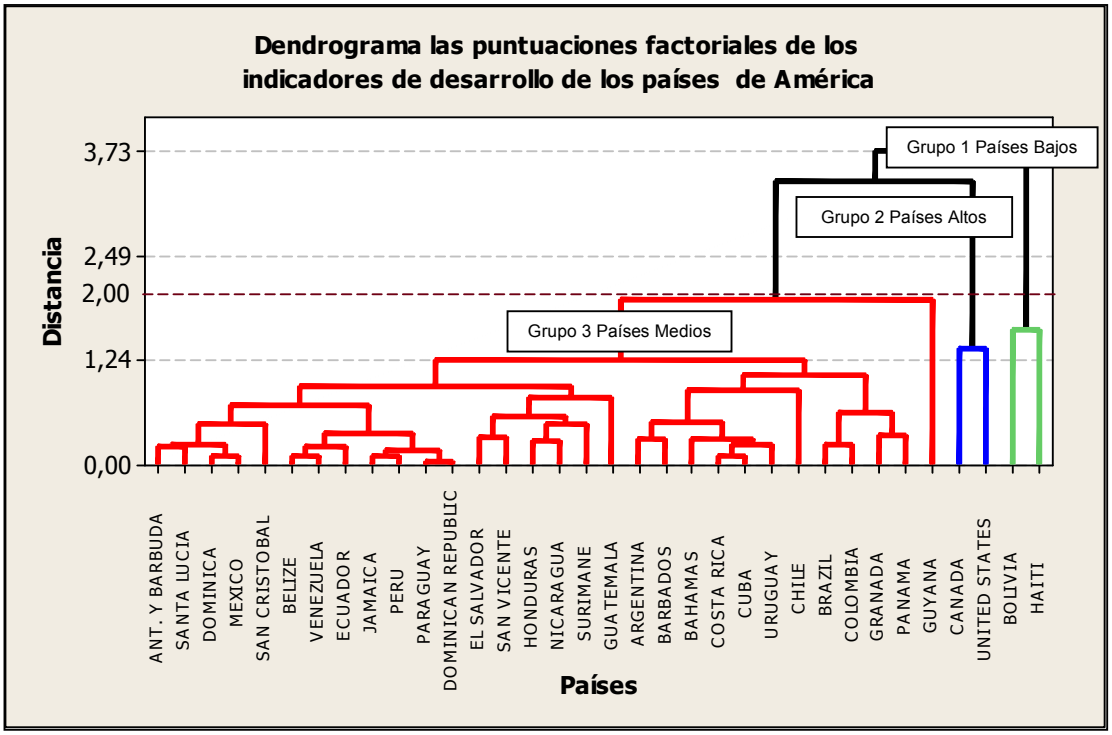
GRÁFICO Nº 06: DENDOGRAMA DE LOS INDICADORES DE DESARROLLO HUMANO DE LOS PAÍSES DE AMÉRICA



Por otro lado, para la confirmación de la agrupación de los países realizado por el Análisis Factorial aplicamos, en el Análisis Cluster, el algoritmo

de Enlace Promedio y Distancia Euclidiana, los resultados muestran que se obtiene 3 grupos (con distancias mayor a 2), definidos según las característica de los países; el conjunto de países Nº 1 está compuesto por Bolivia y Haití, naciones que se caracterizan por un bajo indicador de la esperanza de vida, alta tasa de mortalidad infantil y bajo nivel de gasto en salud pública. Como se esperaba, EE.UU y Canadá se agrupan y forman el cluster Nº 2, ambas naciones presentan altos niveles de desarrollo para la calidad de vida de su población, por otro lado presentan las menores tasas de mortalidad infantil (EE.UU. y Cuba presentan las menores tasas de mortalidad, comparten el puesto Nº 2), a este grupo formado por naciones del norte de América lo denominaremos Países Altos. Los demás países conforman el grupo Nº 3, llamado Países Medios, cabe señalar que Guyana puede separarse, sin embargo por características propias de este país se incluye en el grupo Nº 3. En general, el mapeo de estas agrupaciones (con el Análisis Cluster) confirma los que se formaron en el Análisis Factorial (Gráfico Nº 05). Por otro lado, también muestran una notable diferencia en el desarrollo económico, social y cultural (medidos por el IDH) de los países que subsiste a lo largo del tiempo.

GRÁFICO Nº 07: DENDOGRAMA DE LAS PUNTUACIONES FACTORIALES DE LOS INDICADORES DE DESARROLLO HUMANO DE LOS PAÍSES DE AMÉRICA

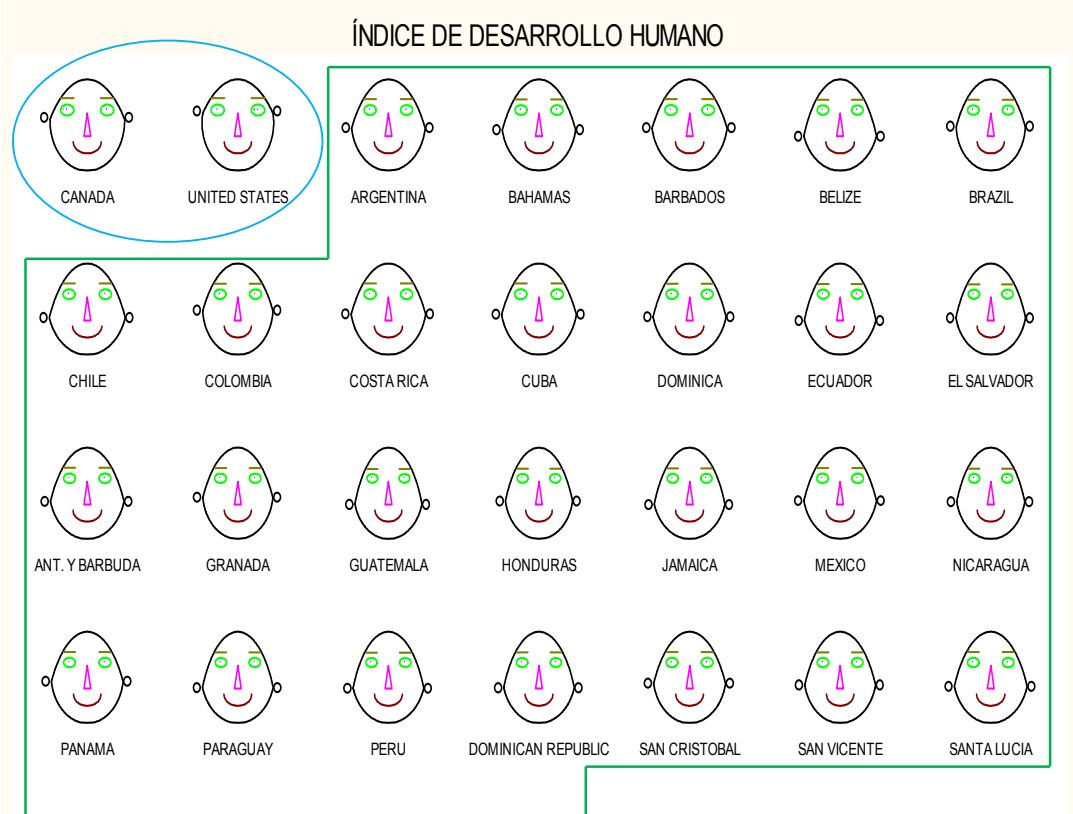


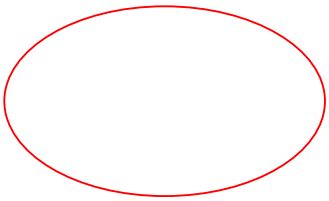
3.8. SEGMENTACIÓN DE LOS PAÍSES AMERICANOS MEDIANTE CARAS DE CHERNOFF

Para una complementación del presente análisis que permita, además, tener una validación adicional de carácter gráfico en cuanto el agrupamiento de los países americanos que se presentó mediante el Análisis factorial y el Análisis Cluster, usamos la técnica de clasificación conocida como "Caras de Chernoff". Este método consiste en una representación de tipo pictórico de los casos (países) usando la magnitud de las variables que se tienen en cuenta en el modelo (admite analizar hasta 12 variables) en correspondencia con los rasgos faciales (Johnson y Wincher, 1982). De tal forma, esta herramienta permite identificar, gráficamente, el grado de similitud entre casos en una cierta proximidad, a partir de la morfología resultante de tales características (variables). Además, el conocimiento de las tendencias de los datos permite clasificarlos de manera visual.

En el siguiente gráfico las Caras de Chernoff representan a los 34 países americanos de acuerdo a sus puntuaciones factoriales obtenidas en el Análisis Factorial, se toma en cuenta solo 2 rasgos: cara y oído.

GRÁFICO Nº 08: CARAS DE CHERNOFF DE LAS PUNTUACIONES FACTORIALES DE LOS INDICADORES DE DESARROLLO HUMANO DE LOS PAÍSES DE AMÉRICA





Fuente: PNUD

Cada rasgo de las caras tiene en cuenta la magnitud relativa para cada país, de la siguiente manera:

- Una cara más ancha expresa una mayor Expectativa de Vida (factor 1 según resultados del Análisis Factorial).
- Mientras más para abajo se encuentre el nivel de las orejas (oídos), menor Calidad de Vida (factor 2 según resultados del Análisis Factorial).

La similitud de las caras indica el agrupamiento de los países. Este comportamiento se observa con el par de países: **Canadá y Estados Unidos**, que muestran caras más anchas que los demás países, lo que indica que su población tiene mayor expectativa de vida, además presentan más arriba la ubicación de las orejas, lo que quiere decir mayor calidad de vida en comparación con la calidad de vida de la población de los demás países. También se alcanza a apreciar rasgos similares en las caras de **Haití y Bolivia**, estos presentan rasgos contrarios al del grupo anterior; caras angostas o delgadas, este rasgo señala que la población de estos países tienen un menor expectativa de vida. Al observar la ubicación de las orejas, estas se encuentran por debajo del centro de la cara que sugiere que la población de este par de países tiene una menor calidad de vida. Los demás países en general muestran similares rasgos; las caras no son muy anchas ni delgadas, lo que indica que la población de estos países presentan una expectativa de vida normal (ni mayor ni menor), en lo que se refiere a la oreja, en general está ubicado en el centro de la cara, lo que advierte una calidad de vida mediana.

En general, las Caras de Chernoff confirman los grupos de países que se formó al aplicar el Análisis Factorial y el Análisis Cluster.

CONCLUSIONES

El Índice de Desarrollo Humano es una medición por país que es usado para clasificarlos en: Desarrollo Humano Alto, Desarrollo Humano Medio y Desarrollo Humano Bajo (ver anexo 2). Sin embargo, no permiten, de forma estricta, establecer una adecuada medición en zonas constituidas, por ejemplo en el continente americano que es el espacio geográfico de análisis de este estudio (además El Caribe). Esto se fundamenta cuando los países de Bolivia y Haití se ubican, según el IDH, en el grupo de los países con Desarrollo Humano Medio (ver anexo 2), cuando en realidad muestran características socio-económicas diferentes a los países americanos que se encuentran en el mismo grupo. De tal forma, sintetizar los indicadores del IDH para luego agrupar a los países, constituye una necesidad en la realización de una adecuada caracterización de países en términos comparativos, principalmente, en el orden internacional. Sin embargo, en América, actualmente las instituciones oficiales no realizan mediciones según zonas constituidas, por ejemplo por continentes.

Precisamente, la realización de un ejercicio para la sintetización de indicadores del IDH y la respectiva agrupación de países, constituye la razón esencial de este trabajo, en cuya realización se han podido precisar los siguientes aspectos principales:

1).- La conclusión general, a partir del ejercicio realizado, es que la agrupación de los países americanos, en términos de su cargas factoriales, contribuye, en buena medida, a una adecuada comparación dentro del contexto americano.

2).- En cuanto a los resultados prácticos se refiere, lo más importante se localiza en el diseño de un método a partir del cual se hace factible el agrupar por zonas constituidas en términos de sus cargas factoriales luego de una adecuada sintetización de los indicadores del IDH.

3).- Las agrupaciones de los países americanos con el Análisis Factorial, Análisis Cluster y Caritas de Chernoff son similares.

4).- Con esta reagrupación se puede tomar decisiones, de gobierno o intergubernamentales, para actuar dentro de cada grupo de países en forma diferente.

5).- Las políticas de ayuda deben de ser diferente para cada grupo de países.

6).- Fortalecer según sea necesario en cada factor hallado (en los indicadores).

BIBLIOGRAFÍA

[1] Programa de las Naciones Unidas para el Desarrollo. *Informe sobre Desarrollo Humano 1995. Nueva York: Oxford University Press. (1995).*

[2] Afifi y Clark, *Computer Aided Multivariate Analysis. Chapman Hall London. (1998).*

[3] C.M. Cuadras*, Métodos de Análisis Multivariante, PPU Barcelona. (1991).*

[4] Hair J.F., Anderson R.E. Tatham R.L. y Black W.C*, Análisis Multivariante Prentice Hall USA. (1999).*

[5] Visauta Vinacua, B. *Análisis Estadístico con SPSS para Windows, McGraw-Hill, Madrid. (1998).*

[6] Anand, Sudhir, y Amartya Sen. *Human Development Index: Methodology and Measurement". Occasional Paper 12. Programa de las Naciones Unidas para el Desarrollo, Oficina encargada del Informe sobre Desarrollo Humano, Nueva York. (1994).*

[7] Pérez López, César. *Técnicas de Análisis Multivariante de Datos, Pearson Eduación S.A., Madrid. (2004).*

[8] Cea D' Ancona, Ángeles. *Análisis Multivariable en la Investigación Social, Editorial Síntesis S.A., Madrid. (2002).*

Páginas web:

http://redalyc.uaemex.mx/redalyc/src/inicio/ArtPdfRed.jsp?iCve=10101404

Visitada el 24/02/2009

http://www.sepeap.org/archivos/libros/estadistica/multivariante/capitulo8.pdf

Visitada el 27/03/2009

http://halweb.uc3m.es/esp/Personal/personas/imolina/MiDocencia/TecnicasInvestigacion0910/SlidesTema4MedidasProximidad.pdf

Visitada el 05/05/2009

ANEXOS

A1. Cálculo del IDH

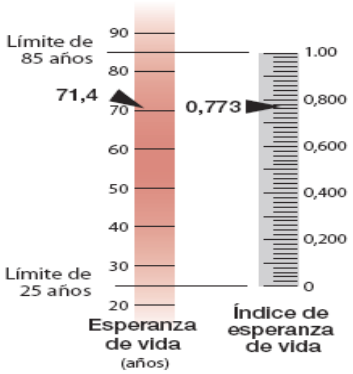
Cómo se calcula el IDH

En este ejemplo se utilizan datos correspondientes a Turquía.

1. Cálculo del índice de esperanza de vida

El índice de esperanza de vida mide los logros relativos de un país en cuanto a esperanza de vida al nacer. Para Turquía, cuya esperanza de vida era de 71,4 años en 2005, el índice de esperanza de vida es de 0,773.

$$\text{Índice de esperanza de vida} = \frac{71{,}4 - 25}{85 - 25} = \mathbf{0{,}773}$$

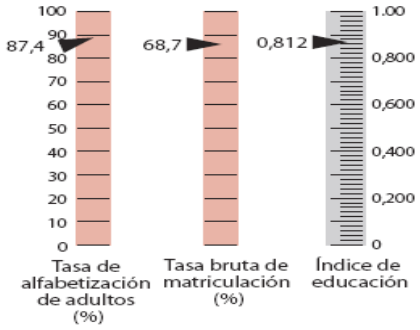


2. Cálculo del índice de educación

El índice de educación mide el progreso relativo de un país en materia de alfabetización de adultos y matriculación bruta combinada en educación primaria, secundaria y terciaria. En primer lugar, se calcula el índice de alfabetización de adultos y el índice de la tasa bruta combinada de matriculación. Luego, se combinan ambos índices para crear el índice de educación, en el que se otorga una ponderación de dos terceras partes a la alfabetización de adultos y de una tercera parte a la tasa bruta combinada de matriculación.Para Turquía, cuya tasa de alfabetización de adultos en 2005 era de 87,4% y la tasa bruta combinada de matriculación era de 68,7% en el mismo año, el índice de educación es de 0,812.

$$\text{Índice de alfabetización de adultos} = \frac{87{,}4 - 0}{100 - 0} = \mathbf{0{,}874}$$

$$\text{Índice bruto de matriculación} = \frac{68{,}7 - 0}{100 - 0} = \mathbf{0{,}687}$$

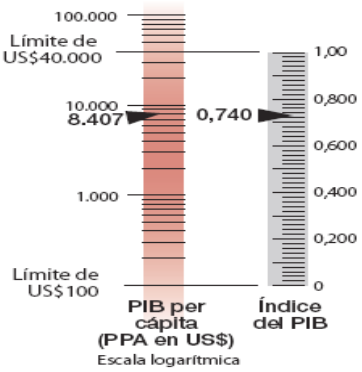


$$\text{Índice de educación} = \text{2/3 (índice de alfabetización de adultos)} + \text{1/3 (índice de matriculación bruta)}$$
$$= \text{2/3 (0,874)} + \text{1/3 (0,687)} = \mathbf{0{,}812}$$

3. Cálculo del índice del PIB

El índice del PIB se calcula utilizando el PIB per cápita ajustado (PPA en US$). En el IDH, los ingresos actúan como sustitutos de todos los demás componentes del desarrollo humano que no se reflejan en una vida larga y saludable ni en la educación adquirida. Los ingresos se ajustan puesto que lograr un nivel respetable de desarrollo humano no requiere ingresos ilimitados. En consecuencia, se utiliza el logaritmo de los ingresos. Para Turquía, cuyo PIB per cápita fue de $8.407 (PPA en US$) en 2005, el índice del PIB es de 0,740.

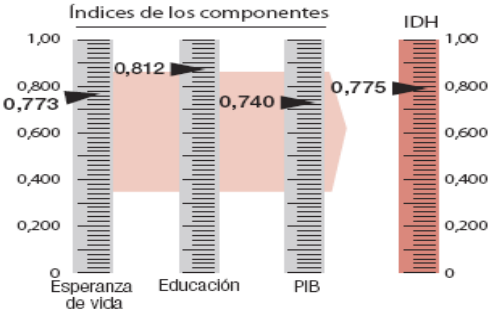
$$\text{Índice del PIB} = \frac{\log(8.407) - \log(100)}{\log(40.000) - \log(100)} = \mathbf{0{,}740}$$



4. Cálculo del IDH

Una vez que se han calculado los índices de cada componente, la determinación del IDH es sencilla. Se trata del promedio simple de los índices de los tres componentes.

$$\text{IDH} = \text{1/3 (índice de esperanza de vida)} + \text{1/3 (índice de educación)} + \text{1/3 (índice del PIB)}$$
$$= \text{1/3 (0,773)} + \text{1/3 (0,812)} + \text{1/3 (0,740)} = \mathbf{0{,}775}$$



A2. Países agrupados según el Índice de Desarrollo Humano

Desarrollo humano alto

(IDH 0,800 y superior)

Albania
Alemania
Antigua y Barbuda
Arabia Saudita
Argentina
Australia
Austria
Bahamas
Bahrein
Barbados
Belarús
Bélgica
Bosnia y Herzegovina
Brasil
Brunei Darussalam
Bulgaria
Canadá
Chile
Chipre
Corea (República de)
Costa Rica
Croacia
Cuba
Dinamarca
Emiratos Árabes Unidos
Eslovaquia
Eslovenia
España
Estados Unidos
Estonia
Finlandia
Francia
Grecia
Hong Kong, China (RAE)
Hungría
Irlanda
Islandia
Israel
Italia
Jamahiriya Árabe Libia
Japón
Kuwait
Letonia
Lituania
Luxemburgo
Macedonia (ERY)
Malasia
Malta
Mauricio
México
Noruega
Nueva Zelandia
Omán
Países Bajos
Panamá
Polonia
Portugal
Qatar
Reino Unido
República Checa
Rumania
Rusia (Federación de)
Saint Kitts y Nevis
Seychelles
Singapur
Suecia
Suiza
Tonga
Trinidad y Tobago
Uruguay
(70 países o territorios)

Desarrollo humano medio

(IDH de 0,500 – 0,799)

Argelia
Armenia
Azerbaiyán
Bangladesh
Belice
Bhután
Bolivia
Botswana
Cabo Verde
Camboya
Camerún
China
Colombia
Comoras
Congo
Djibouti
Dominica
Ecuador
Egipto
El Salvador
Fiji
Filipinas
Gabón
Gambia
Georgia
Ghana
Granada
Guatemala
Guinea Ecuatorial
Guyana
Haití
Honduras
India
Indonesia
Irán (República Islámica del)
Islas Salomón
Jamaica
Jordania
Kazajstán
Kenya
Kirguistán
Lesotho
Líbano
Madagascar
Maldivas
Marruecos
Mauritania
Moldova
Mongolia
Myanmar
Namibia
Nepal
Nicaragua
Pakistán
Papua Nueva Guinea
Paraguay
Perú
República Árabe Siria
República Democrática Popular Lao
República Dominicana
Samoa
San Vicente y las Granadinas
Santa Lucía
Santo Tomé y Príncipe
Sri Lanka
Sudáfrica
Sudán
Suriname
Swazilandia
Tailandia
Tayikistán
Territorios Palestinos Ocupados
Timor-Leste
Togo
Túnez
Turkmenistán
Turquía
Ucrania
Uganda
Uzbekistán
Vanuatu
Venezuela (República Bolivariana de)
Viet Nam
Yemen
Zimbabwe
(85 países o territorios)

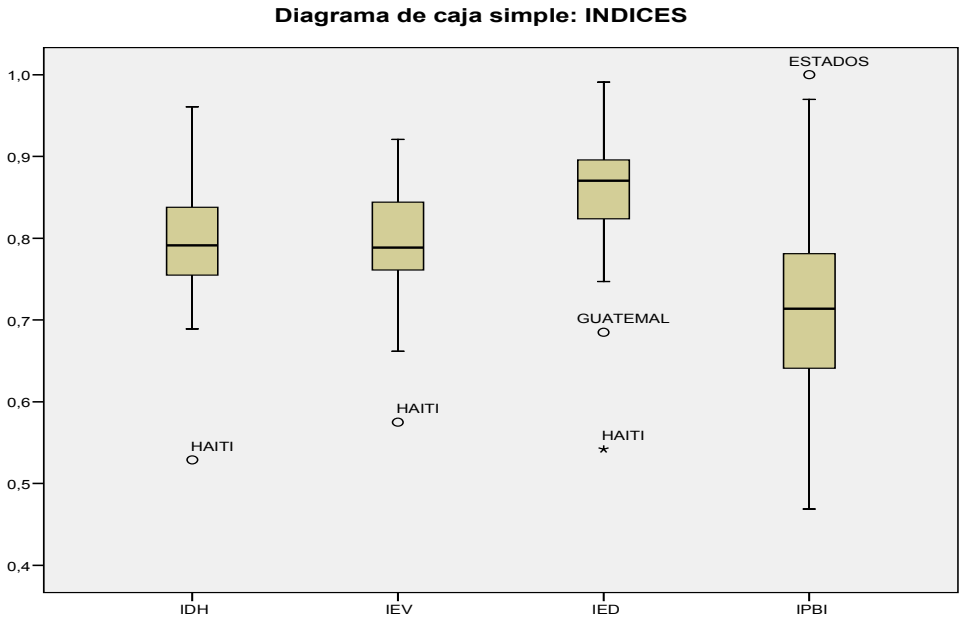
Desarrollo humano bajo

(IDH inferior a 0,500)

Angola
Benín
Burkina Faso
Burundi
Chad
Congo (República Democrática del)
Côte d'Ivoire
Eritrea
Etiopía
Guinea
Guinea-Bissau
Malawi
Malí
Mozambique
Níger
Nigeria
República Centroafricana
Rwanda
Senegal
Sierra Leona
Tanzanía (República Unida de)
Zambia
(22 países o territorios)

Nota: Los siguientes países miembros de la ONU no están incluidos en las listas de países agrupados según el índice de desarrollo humano ya que no se puede calcular su IDH: Afganistán, Andorra, Iraq, Kiribati, Corea (República Popular Democrática de), Liberia, Liechtenstein, Islas Marshall, Micronesia (Estados Federados de), Mónaco, Montenegro, Nauru, Palau, San Marino, Serbia, Somalia y Tuvalu.

A3. Gráfico de cajas: Algunos indicadores del IDH de los países de América



A4. Valores propios

Valores propios:

	1	2	3	4	5	6	7
Valor propio	4.248	1.311	0.549	0.451	0.233	0.111	0.097
Variabilidad (%)	60.686	18.729	7.842	6.437	3.330	1.592	1.383
% acumulado	60.686	79.415	87.258	93.695	97.025	98.617	100.000

A5. Vectores propios

Vectores propios:

	F1	F2	F3	F4	F5	F6	F7
Índice de la Esperanza de Vida	0.423	-0.254	-0.020	0.130	-0.743	-0.032	0.432
Tasa mortalidad infantil < de 1 año	-0.356	0.442	-0.323	-0.374	-0.556	0.267	-0.230
Tasa mortalidad infantil < de 5 años	-0.423	0.348	0.169	-0.033	0.104	-0.070	0.809
Índice del PBI	0.440	0.117	0.245	-0.352	0.199	0.739	0.152
Gasto público en salud en % del PBI	0.261	0.585	-0.249	0.706	0.056	0.161	-0.027
Gasto en salud per cápita en PPA en $	0.326	0.500	0.548	-0.227	-0.131	-0.488	-0.194
Índice de educación	0.385	0.117	-0.666	-0.411	0.262	-0.335	0.211

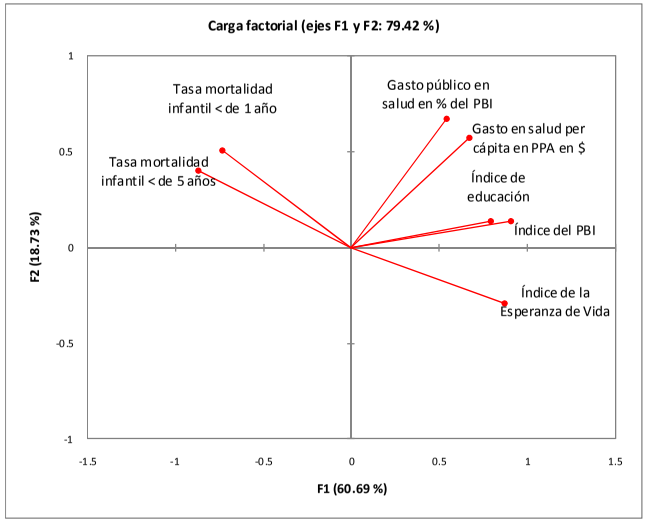
A6. Coordenadas factoriales

Coordenadas factoriales sin rotación:

	F1	F2	comunalidad inicial	Comunalidad final	Varianza específica
Índice de la Esperanza de Vida	0.872	-0.290	1.000	0.845	0.155
Tasa mortalidad infantil < de 1 año	-0.733	0.506	1.000	0.794	0.206
Tasa mortalidad infantil < de 5 años	-0.871	0.399	1.000	0.917	0.083

Índice del PBI	0.906	0.134	1.000	0.839	0.161
Gasto público en salud en % del PBI	0.538	0.670	1.000	0.738	0.262
Gasto en salud per cápita en PPA en $	0.671	0.572	1.000	0.778	0.222
Índice de educación	0.793	0.134	1.000	0.647	0.353

A7. Gráfico de cargas factoriales sin rotación

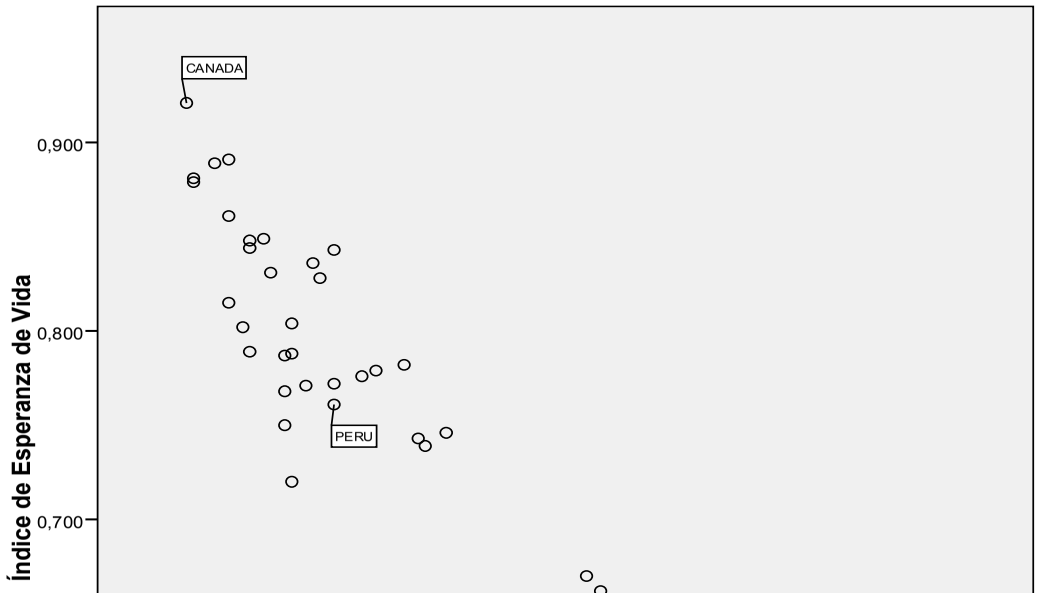


A8. Matriz de correlaciones

Matriz de correlaciones(a)

		Índice de Esperanza de Vida	Índice de Educación	Índice del PBI	Gasto en salud publica, % del PBI	Gasto en salud, per capita en dolares	Tasa de mortalidad infantil (por cada 1000 nacidos vivos)	Tasa de mortalidad infantil menores de 5 años(por cada 1000 nacidos vivos)
Correlación	Índice de Esperanza de Vida	1	0.601	0.697	0.307	0.417	-0.720	-0.863
	Índice de Educación	0.601	1	0.700	0.474	0.457	-0.375	-0.668
	Índice del PBI	0.697	0.700	1	0.447	0.746	-0.588	-0.696
	Gasto en salud publica, % del PBI	0.307	0.474	0.447	1	0.587	-0.132	79-0.237
	Gasto en salud, per capita en dolares	0.417	0.457	0.746	0.587	1	-0.255	-0.317
	Tasa de mortalidad infantil (por cada 1000 nacidos vivos)	-0.720	-0.375	-0.588	-0.132	-0.255	1	0.783
	Tasa de mortalidad infantil menores de 5 años(por cada 1000 nacidos vivos)	-0.863	-0.668	-0.696	-0.237	-0.317	0.783	1
Sig. (Unilateral)	Índice de Esperanza de Vida		0.000	0.000	0.038	0.007	0.000	0.000
	Índice de Educación	0.000		0.000	0.002	0.003	0.014	0.000

A9. Relación enre las variables Tasa de mortalidad infantil menores de 5 años (por cada 1000 nacidos vivos) con el Índice de Esperanza de Vida



A10. Matriz ANTI-IMAGEN

Matrices anti-imagen

		i_esp_vida_9	i_educacion_10	i_pbi_11	Gasto en salud publica, % del PBI	Gasto en salud, per capita en dolares	Tasa de mortalidad infantil (por cada 1000 nacidos vivos)	Tasa de mortalidad infantil menores de 5 años(por cada 1000 nacidos vivos)
Covarianza anti-image	i_esp_vida_9	.226	.009	-.010	-.029	-.030	.032	.102
	i_educacion_10	.009	.334	-.093	-.142	.049	-.125	.090
	i_pbi_11	-.010	-.093	.174	.049	-.158	.059	.024
	Gasto en salud publica, % del PBI	-.029	-.142	.049	.572	-.172	.009	-.025
	Gasto en salud, per capita en dolares	-.030	.049	-.158	-.172	.288	-.035	-.045
	Tasa de mortalidad infantil (por cada 1000 nacidos vivos)	.032	-.125	.059	.009	-.035	.313	-.100
	Tasa de mortalidad infantil menores de 5 años(por cada 1 nacidos vivos)	.102	.090	.024	-.025	-.045	-.100	.142
Correlación anti-image	i_esp_vida_9	.868[a]	.034	-.051	-.080	-.117	.120	.568
	i_educacion_10	.034	.757[a]	-.385	-.326	.159	-.385	.415
	i_pbi_11	-.051	-.385	.771[a]	.154	-.705	.254	.154
	Gasto en salud publica, % del PBI	-.080	-.326	.154	.744[a]	-.423	.022	-.087
	Gasto en salud, per capita en dolares	-.117	.159	-.705	-.423	.651[a]	-.117	-.220
	Tasa de mortalidad infantil (por cada 1000 nacidos vivos)	.120	-.385	.254	.022	-.117	.784[a]	-.475
	Tasa de mortalidad infantil menores de 5 años(por cada 1 nacidos vivos)	.568	.415	.154	-.087	-.220	-.475	.753[a]

a. Medida de adecuación muestral

A11. Puntuaciones o cargas factoriales

Puntuaciones factoriales de las observaciones (rotación Varimax)

País	F1	F2
ANT. Y BARBUDA	0,659	-0,372
ARGENTINA	0,495	0,627
BAHAMAS	0,492	0,252
BARBADOS	0,798	0,641
BELIZE	0,642	-0,955
BOLIVIA	-3,031	0,870
BRAZIL	-0,448	0,850
CANADA	0,787	2,446
CHILE	1,176	-0,378
COLOMBIA	-0,230	0,891
COSTA RICA	0,728	0,265
CUBA	0,800	0,202
DOMINICA	0,528	-0,297
ECUADOR	0,415	-1,021
EL SALVADOR	-0,128	-0,571
UNITED STATES	0,235	3,716
GRANADA	-0,218	0,349
GUATEMALA	-0,537	-1,106
GUYANA	-1,448	0,542
HAITI	-3,605	-0,614
HONDURAS	-0,713	-0,436
JAMAICA	0,241	-1,005
MEXICO	0,437	-0,271
NICARAGUA	-0,475	-0,549
PANAMA	0,127	0,402
PARAGUAY	0,179	-0,821
PERU	0,210	-0,932
DOMINICAN REPUBLIC	0,155	-0,840
SAN CRISTOBAL	0,238	0,001
SAN VICENTE	0,071	-0,323
SANTA LUCIA	0,553	-0,538
SURIMANE	-0,406	-0,083
URUGUAY	0,725	0,012
VENEZUELA	0,550	-0,952